Integrating Materials and Manufacturing Innovation
https://doi.org/10.1007/s40192-022-00267-2

TECHNICAL ARTICLE

Feature Engineering for Microstructure–Property Mapping in Organic Photovoltaics

Sepideh Hashemi[1] · Baskar Ganapathysubramanian[2] · Stephen Casey[3] · Ji Su[4] · Surya R. Kalidindi[1,2]

Received: 21 February 2022 / Accepted: 10 June 2022

Abstract
Linking the highly complex morphology of organic photovoltaic (OPV) thin films to their charge transport properties is critical for achieving high performance material systems that facilitate cost-efficient energy harvesting. In this paper, the current Materials Knowledge Systems (MKS) framework was extended so that it was able to establish reduced-order high-fidelity structure–property linkages for OPV films. Specifically, the following extensions were needed: (i) the proper application of digital image processing algorithms to identify the salient local material states in OPV microstructures controlling the charge transport phenomenon, (ii) computationally efficient feature engineering that not only utilized 2-point spatial correlations and principal component analysis, but also two new distance-based metrics, and (iii) the successful application of a localized version of the Gaussian process (laGP) together with an active learning Cohn (ALC) for building the desired surrogate models linking the OPV microstructures to their short-circuit currents. It is demonstrated that the extended MKS framework can produce high-fidelity structure–property linkages for OPV films.

Keywords Unsupervised feature engineering · Reduced-order models · Structure–property linkages · Organic photovoltaics · Charge transport · Gaussian processes

Introduction

Flexible, lightweight, and wearable solar cells offer a promising solution to cheap energy harvesting for consumer products as well as residential applications. Over the past decade, rapid developments in synthetic chemistry have resulted in organic photovoltaic systems that have pushed single-junction organic photovoltaic (OPV) efficiencies over 16%. These novel materials—electron-donors and electron-acceptors—provide tremendous opportunities for improved performance, reaching the performance of silicon-based photovoltaics. In conjunction with synthesis advances, a large body of work has demonstrated that the microstructure in the active layer is key to high performance devices. Thus, tailoring the morphology in the active layer of OPVs continues to be crucial for maximizing performance. More importantly, advances in self-assembly suggests the possibility of remarkable control of the active layer morphology.

Despite the importance of morphology to OPV device performance, it remains a challenge to comprehensively and rapidly map morphologies to performance. The availability of reliable and fast structure–property models could enable domain scientists to (a) explore, identify and design "ideal" morphologies that maximize performance, (b) identify microstructure features that positively (or negatively) impact performance, and (c) quantify how perturbations to the morphology (due to oxidation, annealing or aging) degrade performance.

Past approaches of investigating structure–property linkages relied on full-physics simulators either discrete (kinetic) Monte Carlo models, or continuum drift–diffusion models. These models are typically expensive to deploy, and sequential deployment for exploration or optimization has been shown to be prohibitively expensive. Similarly, rapid design exploration using such full-physics simulators is

✉ Surya R. Kalidindi
surya.kalidindi@me.gatech.edu

1 George W, Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

2 Department of Mechanical Engineering, Iowa State University, Ames, IA 50011, USA

3 NASA Langley Research Center, Hampton, VA 23681, USA

4 School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

Published online: 18 July 2022

typically not possible, even with access to high performance computing resources.

Recent approaches overcome this challenge by first creating a diverse dataset of annotated morphologies and their performances, and then utilizing data-driven tools on this dataset to construct *low-computational cost surrogate structure–property models*. Such a strategy amortizes the cost of creating a large, annotated dataset across multiple studies. Additionally, property annotation on this dataset using the full-physics simulators are embarrassingly parallel, thus, optimally utilizing HPC resources.

Such structure–property surrogate models—especially in the context of OPV—have been successfully constructed and deployed for design optimization, process–structure–property linkages, sensitivity analysis, and other studies. However, most of these studies have:

- either relied on manual 'featurization' of the morphologies based on knowledge of the photophysics [1–3]. While very useful, such approaches are non-trivial and generally time-consuming. Additionally, manual featurization carries the risk of overlooking or neglecting important features,
- or utilized the full raw morphology data to construct structure–property linkages [4]. However, these approaches need massive datasets to train good surrogate models due to the large input dimensionality (of the morphology image). Additionally, the resultant surrogates are complex and usually not interpretable.

In this work, we bridge these two extremes by using a principled approach of unsupervised featurization of the morphologies. These low dimensional set of features are then used to train an accurate structure–property surrogate model. Specifically, the recently developed Material Knowledge System (MKS) framework [5–9] offers a data-driven framework for unsupervised feature engineering of material microstructures. This framework employs a voxelized representation of microstructures to efficiently compute the 2-point spatial correlations [10–12] and perform principal component analysis (PCA) [13, 14] on them to identify a sufficiently small number of features representing the complex material microstructure. The feature engineering developed in the MKS framework is unsupervised in that the microstructure feature selection is completely uninfluenced by the output variables targeted by the surrogate model. Although a large number of options exist for building the surrogate models of interest, recent work in the MKS framework [8, 9, 15–20] has demonstrated that Gaussian process regression (GPR) [14, 21] offers advantages because of its ability to formulate nonparametric models while allowing for a rigorous consideration of the prediction uncertainty.

Certain extensions are needed to the current MKS framework in order to apply it successfully to the present problem: (i) A large number of pixel-scale (local) material states need to be considered, which is expected to be significantly larger than those encountered in prior case studies. This is because of the need to consider not only the donor and acceptor pixels, but also the different types of the donor–acceptor interfaces present in the microstructure. This challenge is addressed in this work by applying digital image processing techniques in combination with the previously established MKS feature selection methods that utilized fast-Fourier transform (FFT) based computations of 2-point spatial correlations and PCA. (ii) The small thickness of the films necessitated the development and use of new features that specifically accounted for the distances of the different types of the donor–acceptor interfaces from the top and the bottom surfaces of the films [22]. (iii) Given the large dimensional and large input domains involved, a localized version of the Gaussian process (laGP) is needed. In this work, the laGP is implemented together with an active learning Cohn (ALC) for building the desired surrogate models. This paper demonstrates that this extended MKS framework is capable of producing highly accurate reduced-order models linking the OPV microstructures to their short-circuit currents.

Background

Microstructure and Photovoltaic Property Dataset

We utilize a curated dataset of microstructure images created by solving the Cahn–Hilliard equation [23] with varying initial conditions. The Cahn–Hilliard equation [23] describes phase separation occurring in a binary mixture and has been shown to be a good representation of morphology evolution during fabrication of organic blend thin films [24–26] that are the typical active layer in OPV's. The image data arising from these simulations provide a rich dataset for constructing structure–property surrogate models [4]. The dataset is a collection of 33,552 microstructure images of 101×101 pixels in resolution. Each image is grayscale, with the value of each pixel ranging between 0 and 1.

Each microstructure is virtually interrogated to extract its current–voltage characteristics, by solving a morphology aware (i.e., spatially heterogeneous) photophysics device model. We deploy a validated, in-house software that uses a finite element based solution strategy for solving the photophysics device model [27–29]. The photophysics model is described by the steady state *excitonic drift diffusion (XDD) equations*. The XDD equations are a set of four tightly coupled partial differential equations that model the optoelectronic physics of energy harvesting in organic photovoltaic

devices. The photophysics consists of the following stages (also illustrated in Fig. 1):

- Incident solar radiation causes the generation of energetically active electron–hole pairs, called excitons (denoted by χ), in the donor regions of the microstructure. These excitons diffuse across the microstructure and have a finite lifetime before becoming ground state electron–hole pairs;
- Excitons that diffuse and reach the donor–acceptor interface undergo dissociation into electrons (denoted by N) and holes (denoted by P) at the donor–acceptor interface. The dissociation mechanism is material and field dependent (denoted by D);
- These generated charges (N,P) traverse the microstructure and reach their corresponding electrodes (cathode and anode) to produce a current. Two mechanisms are responsible for driving carrier transport or current flow. First, the drift, which is caused by the presence of an electric field (denoted as the gradient of the potential, $\nabla\varphi$), and second, the diffusion, which is caused by a spatial gradient of electron or hole concentration;
- The distribution of electrons and holes in the microstructure interacts with the applied voltage and influences the electrostatic potential φ across the microstructure. Finally, electrons and holes can recombine (denoted by ρ) to create excitons

The photophysics described above is encoded using the exciton drift diffusion (XDD) equations [27]. In prior work, these XDD equations were solved to get the performance of the OPV device, which is characterized by the short-circuit current J_{sc}. XDD simulation results for each of the 33,552 microstructures generated earlier provide us the photophysics properties (J_{sc}).

Feature Engineering Using MKS Framework

In the MKS framework, the uniformly discretized (i.e., voxelated) representative volume elements (RVEs) of the material microstructures are denoted by an array, m_s^h, whose elements denote the volume fractions of the material state h found at voxel s. Microstructural domains where each voxel is occupied fully by a specific material state lead to microstructure arrays where the value of m_s^h is either 0 or 1. Although it may be tempting to use m_s^h directly as the feature set, it should be recognized that it lacks translational invariance. The MKS framework employs the framework of 2-point spatial correlations [10–12], which are essentially auto- and cross-correlations of material state maps of the microstructure. Mathematically, the discretized set of 2-point spatial correlations, denoted as $f_r^{hh'}$, are computed as

$$f_r^{hh'} = \frac{1}{S_r}\sum_s m_s^h m_{s+r}^{h'} \tag{1}$$

where h and h' index all of the material states present in the studied material system, $\boldsymbol{r}$ indexes a set of discretized vectors arising from the voxelization used to define m_s^h, and S_r denotes the total number of pixels that allow for placement of vectors $\boldsymbol{r}$ within the microstructural domain. The computations implied in Eq. (1) can be efficiently carried out using the fast Fourier transform (FFT) algorithm [30, 31].

The complete set of 2-point spatial correlations computed using Eq. (1) produces a large unwieldy set of features. In the MKS framework, a smaller set of salient features is identified (i.e., feature engineering) by performing principal component analysis (PCA) [13, 14], which (rotationally) transforms the data into a new space where the axes are organized by their ability to account for the variance in the dataset. The new orthogonal axes and the new coordinates obtained from the PCA are then referred to as PC scores

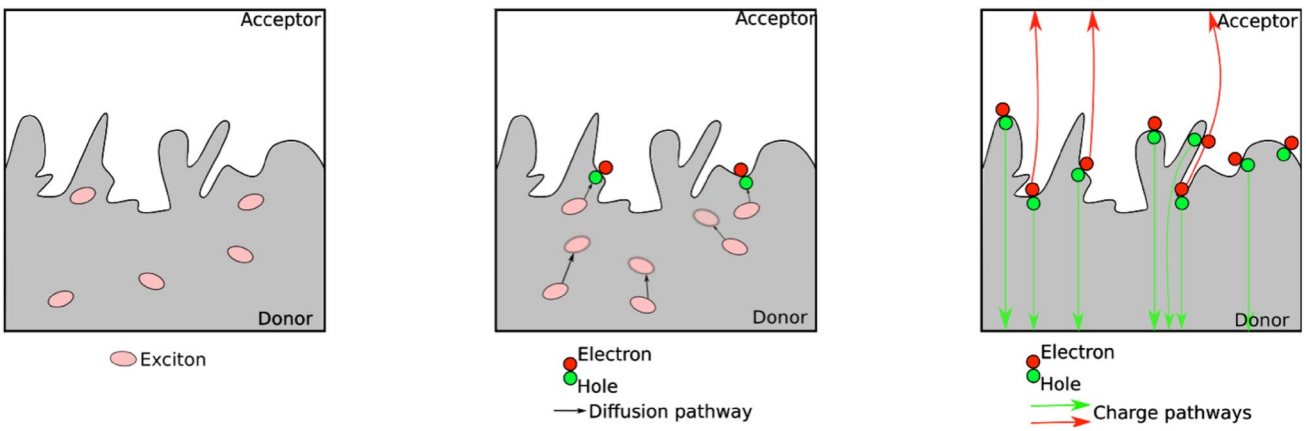

Fig. 1 Schematic illustrating the various stages of the photophysics process (see main text for detailed description)

and PC basis, respectively. Prior studies have often shown a drastic dimensionality reduction going from $\sim 10^5 - 10^6$ original microstructural features to less than $\sim 10 - 15$ PCs [6, 8, 15, 32, 33].

Gaussian Process Regression Models

Although many surrogate model building approaches can be used for building structure–property linkages, prior work has shown the benefits of using Gaussian process regression (GPR) in combination with the MKS feature engineering described earlier [8, 9, 15–20]. GPR is particularly powerful when building surrogate models for complex nonlinear systems/phenomena, where the parametric model forms are not yet established. The other main advantage of GPR lies in the quantification of the uncertainty associated with the model predictions.

In the GPR-MKS framework, the reduced-order structure–property linkage of interest can be decomposed into a linear mean function m and an error function ε often modeled as a zero-mean Gaussian process. Mathematically, the desired model is expressed as [21]

$$p = m(\boldsymbol{\gamma}) + \varepsilon \tag{2}$$

$$m(\boldsymbol{\gamma}) = \beta_0 + \sum_{i=1}^{R} \beta_i \gamma_i \tag{3}$$

$$\varepsilon \sim \mathcal{GP}\left(0, k\left(\boldsymbol{\gamma}, \boldsymbol{\gamma}'\right)\right) \tag{4}$$

where p is the target property (i.e., output), $\boldsymbol{\gamma}$ is the input feature vector consisting of R PCs, $\boldsymbol{\beta}$ are coefficients of the linear model, and $k\left(\boldsymbol{\gamma}, \boldsymbol{\gamma}'\right)$ is the GP's covariance function. The automatic relevance determination squared exponential (ARD-SE) kernel [21] has often been used to define the GP's covariance. The ARD-SE kernel is mathematically expressed as

$$k\left(\boldsymbol{\gamma}, \boldsymbol{\gamma}'\right) = \sigma_f^2 \exp\left[-\frac{1}{2}\sum_{l=1}^{R}\frac{\left(\gamma_l - \gamma_l'\right)^2}{\sigma_l^2}\right] + \sigma_n^2 \delta_{\boldsymbol{\gamma}\boldsymbol{\gamma}'} \tag{5}$$

where the scaling factor σ_f, length scale σ_l, and noise factor σ_n are hyperparameters of the kernel function, and $\delta_{\boldsymbol{\gamma}\boldsymbol{\gamma}'}$ is the Kronecker delta. The hyperparameter σ_n determines the homoscedastic noise in the target predictions. The hyper parameter σ_f controls the amplitude of the variance in the output. The length scale σ_l automatically determines the relevance of input features on the predictions. Higher values of σ_l results in smoother predictions, indicating minimal influence on the output prediction. The values of hyperparameters need to be optimized during the model building process to obtain the best model.

The joint distribution of the observed training data ($\boldsymbol{X}$) and the unobserved test data ($\boldsymbol{X}_*$) is given by [21]

$$\begin{bmatrix} \boldsymbol{p} \\ \boldsymbol{p}_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(\boldsymbol{X}, \boldsymbol{X}) & K_*\left(\boldsymbol{X}, \boldsymbol{X}_*\right) \\ K_*^{\dagger}\left(\boldsymbol{X}, \boldsymbol{X}_*\right) & K_{**}\left(\boldsymbol{X}_*, \boldsymbol{X}_*\right) \end{bmatrix}\right) \tag{6}$$

The predictive posterior is obtained from conditioning the joint distribution fully defined by its mean and covariance [21]:

$$\boldsymbol{\mu}_* = K_*^{\dagger} K^{-1} \boldsymbol{p} \tag{7}$$

$$\boldsymbol{\Sigma}_* = K_{**} - K_*^{\dagger} K^{-1} K_*$$

The main computationally intensive operation in GP formulation is the inversion of the kernel matrix which scales as $O\left(N^3\right)$. Although this is a one-time computation, in case of large ensemble of training data, the computation and storage of K^{-1} present significant challenges. Prior studies have addressed these challenges using methods such as low-rank approximations to GPs [21, 34], treed GPs [35, 36] and local approximate GP (laGP) [37, 38]. Recent research has demonstrated that low-rank approximations and treed GPs tend to over-smooth the data, might impose an upper limit on the data size and typically take longer to compute [39]. The recently developed laGP model is particularly attractive as it scales well with the data size, allows for non-stationarity modeling, and is highly parallelizable. The laGP model is a local variant of the GPR which employs a local subset of the data to train separate GPs for each target point. The subset of data can be chosen as n nearest neighbors of the target point. However, this simple criterion does not yield the optimum predictions. Instead, the laGP approach utilized in this work employs the active learning Cohn (ALC) method [38, 40] to sequentially update the chosen subset of the training points. The ALC method sequentially identifies points whose addition to the local subset maximizes the expected information gain by maximizing the reduction in the prediction variance. More specifically, for each prediction, the first n_0 nearest neighbors to the target point are chosen as the initial set for constructing the first laGP model. Then the ALC method is applied over all of the remaining points to identify the new point to be added to the next update of the laGP model. Points are sequentially identified until no further improvement to model is observed. Thus, the size of the final set of neighbors utilized in each laGP model is represented by $n_d = n_0 + n_{ALC}$, where n_{ALC} is the total number of points selected by the application of the ALC method. Since the number of neighbors selected is typically quite small ($n_d \ll N$), laGP successfully circumvents the aforementioned challenges in the use of global GPR on large datasets.

Microstructure–Property Models for Photovoltaic Polymers

The workflow used in this paper for building the surrogate microstructure–property models for OPVs will involve two main steps: (i) unsupervised feature engineering of the microstructure using the MKS framework, and (ii) establishing the laGP models using the engineered features. Further details of these steps are described next.

Material States in OPV Microstructures

The grayscale OPV microstructures (with each pixel value ranging between zero and one) obtained from solving the Cahn–Hilliard equation (summarized in "Microstructure and Photovoltaic Property Dataset" section) are thresholded into binary microstructures consisting of donor (D) and acceptor (A) phases (i.e., material state binerization). In this study, a threshold of 0.5 was used to convert the gray-scale microstructures into binary microstructures. The charge transport of OPV materials is affected by the shape, size, spacing distribution of donor and acceptor regions as well as their connectivity to their corresponding electrodes. More specifically, in order for the OPV microstructures to exhibit efficient charge transport, the donor and the acceptor regions should be directly connected to the corresponding electrodes positioned at top and bottom surfaces of the thin films, respectively. In other words, the donor/acceptor pixels connected to their respective electrodes are expected to be very productive, while those not connected to their respective electrodes are expected to be fairly non-productive. Therefore, it was decided to define four different material local states for labeling the individual pixels in the microstructures: (i) D^{Λ}—donor pixels connected to the top surface, (ii) D°—donor pixels unconnected to the top surface, (iii) $A^{\vee}$—acceptor pixels connected to the bottom surface, and (iv) A°—acceptor pixels unconnected to the bottom surface. These constitute the first set of material local states identified for this work.

In addition, the different types of the donor–acceptor interfaces present in the microstructure affect the charge transport in very different ways. Any interface pixels between two connected regions (i.e., regions connected to their respective electrodes), defined as $I_1 = \left(D^{\Lambda}, A^{\vee}\right)$, are expected to contribute the most to the charge transport. It can also be seen that any interface pixels between two unconnected regions, $I_2 = (D^{\circ}, A^{\circ})$, are fairly non-productive. The other two sets of interface pixels, $I_3 = \left(D^{\Lambda}, A^{\circ}\right)$ and $I_4 = \left(D^{\circ}, A^{\vee}\right)$, are considered semi-effective. The four sets of interfaces thus defined constitute the second set of identified material local states.

As a final consideration, the charges created in OPV microstructures typically move through the donor and acceptor regions that are directly connected to the top and bottom electrodes (D^{Λ} and $A^{\vee}$), respectively. In addition, if unconnected donor/acceptor regions (D° and A°) are considerably close to their respective electrodes, they can also play an important role in the charge transport [22], especially in microstructures that comprised only unconnected donor/acceptor regions. Note that the charge transport in such microstructures is inversely related to the distance of the closest D° and A° from their relevant electrodes (i.e., shorter the distance, higher the charge transport). These insights were used to define two additional distance-based metrics described later.

In order to properly account for all of the physical insights described above, we devised and implemented a 3-step procedure to assign material local states to each pixel in each OPV microstructure. In the first step, we assign one of the four material local states described above to each voxel in the OPV microstructure: D^{Λ}, D°, $A^{\vee}$, and A° (see Fig. 2a). This was achieved by first considering the donor phase as the foreground (i.e., assigning values of one to donor pixels and zero to acceptor pixels) and using a cluster labeling algorithm [41] to identify uniquely the connected sets of the donor pixels (i.e., donor clusters). The pixels in the donor clusters connected to the top surface were all assigned the material state D^{Λ}, while the rest of the donor pixels were assigned the material state D°. A similar procedure was performed to assign the material states $A^{\vee}$ and A°. Note that the assignment of these four material states is mutually exclusive. In other words, every pixel in the microstructure is assigned only one of the four material states mentioned above.

In the second step, we have defined an additional material local state identifying the different types of interfaces between the donor and acceptor pixels. This additional material state is assigned only to the interface pixels. As already described, a total of four different interfaces are possible: (D^{Λ}, $A^{\vee}$), (D^{Λ}, A°), (D°, $A^{\vee}$), and (D°, A°) (see the microstructure shown in Fig. 2b). In this work, we adopted a 2-pixel interfacial region that included the first pixel on either side of the interface. The interface pixels are identified using a computational strategy developed in prior work for 2-phase microstructures [8]. This method is used due to its computational efficiency, derived from the use of the convolution kernel shown in Fig. 2b on selected foreground material states. The result of this computations is an integer $c_i \in [1 : 4]$ for each pixel i, which identifies the four desired classes of interfacial pixels described earlier (note that the interior pixels within the foreground and background would exhibit values zero and five, respectively). In this work, special considerations were made to account for the non-periodicity of the microstructures. Specifically, this challenge

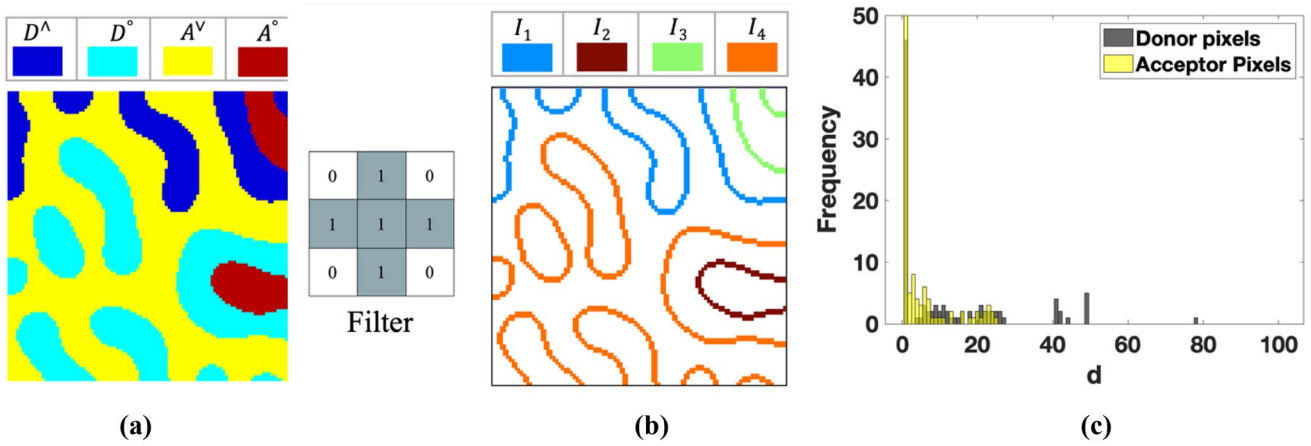

Fig. 2 Labeling of the material local states to each pixel of a selected OPV microstructure. **a** Each pixel is assigned one of the four material local states corresponding to connected/unconnected donor/acceptor pixels. Connectivity in this context refers to whether the donor/acceptor pixels are connected to their corresponding electrodes at the top/bottom surfaces. **b** Each interface pixel is assigned one of the four interfaces. The interfacial region is considered to be 2-pixel thick, comprising both pixels on either side of the interface. The convolution kernel used to identify the interfaces is shown on the right. **c** A third material state is assigned to the top and bottom rows of pixels based on the shortest distance, d, of the donor (acceptor) pixels from the top (bottom) surface. The plot shows distribution of d for the selected microstructure

was addressed using suitable zero-padding schemes [31]. For non-periodic microstructures, the sets of edge pixels and corner pixels were identified separately; edge pixels with $c_i \in [1:3]$ and corner pixels with $c_i \in [1,2]$ denote interfacial pixels. By applying the procedure described above to each phase (i.e., treating each phase as foreground one at a time), each interface pixel can be mapped uniquely to one the aforementioned four types of interfaces. Figure 2b shows the labeling of the interface pixels for the example microstructure shown in Fig. 2a.

In the last step of the unsupervised feature identification procedure employed in this study, we identify a third local material state descriptor, which is applied only for the top/bottom rows of pixels connecting to the electrodes. This feature is designed to capture the effect arising from the shortest distance of donor/acceptor pixels from their respective electrodes, which essentially reflects the transport distance for the generated charges to complete the circuit. As already mentioned, this feature is especially important for microstructures where the donor/acceptor pixels are not in direct contact with their corresponding electrodes. Figure 2c presents a histogram of the shortest vertical distance of the donor (acceptor) pixel to the top (bottom) surface, d, for the example microstructure shown in Fig. 2a. For our work, it is necessary to suitably scale these distances to reflect the fact that larger values of d do not contribute significantly. It was decided to use $\exp(-d/\lambda)$ as the feature value for each top/bottom pixel, with $\lambda = 10$nm reflecting the expected diffusion length for charge transport [22, 42]. Consequently, the feature value is one when the pixels are in direct contact and exponentially decreases when there is a gap. The rate of decrease is controlled by the value of λ, i.e., pixels farther than λ are assumed to make fairly insignificant contributions to the charge transport.

After labeling the material local states, the next step involves the computation of the important microstructure statistics. The central challenge comes from the large number of spatial statistics that could be computed. In the present case, since there are a total of eight material local states (four acceptor/donor states and four interface states), one can potentially define a total of $8^2 = 64$ sets of spatial correlations (including auto-correlations and cross-correlations). Since each set of spatial correlations has a total of $101 \times 101 = 10,201$ features, the full set of features becomes unwieldy for establishing surrogate models. In prior work [8] on correlating the effective permeability of a porous solid to its pore structure, it was observed that the auto-correlations of the material local states (including interface states) were adequate for producing high fidelity structure–property linkages. Utilizing the insights from that work, we have included only the following sets of spatial correlations in establishing the surrogate models presented in this work: i) 2-point spatial auto-correlations for each of the four main material local states $\left\{f_{\boldsymbol{r}}^{D^{\wedge}D^{\wedge}}, f_{\boldsymbol{r}}^{D^{\circ}D^{\circ}}, f_{\boldsymbol{r}}^{A^{\vee}A^{\vee}}, f_{\boldsymbol{r}}^{A^{\circ}A^{\circ}}\right\}$, and ii) 2-point spatial auto-correlations for each of the four interfacial local states $\left\{f_{\boldsymbol{r}}^{I_1I_1}, f_{\boldsymbol{r}}^{I_2I_2}, f_{\boldsymbol{r}}^{I_3I_3}, f_{\boldsymbol{r}}^{I_4I_4}\right\}$. Even using only this subset of spatial correlations produces a total of $8 \times 10,201 = 81,608$ features. As already described in "Feature engineering using MKS framework"Section, PCA is applied to obtain a small number of features (i.e., PC scores) as inputs to the surrogate structure–property models. Prior to application of PCA, each of the eight sets of spatial

correlations is scaled to exhibit the same variance across the entire dataset. This is necessary due to the fact that PCA aims to capture the variance in the dataset in the smallest number of terms. Therefore, scaling the different sets of spatial correlations ensures that each set of spatial correlations is equally weighted in the PC representations. In this work, for reasons already explained, the averaged values of $\exp(-d/10)$ for both electrodes, denoted as $\left\{\delta^{\Lambda}, \delta^{\vee}\right\}$, are used as additional features (i.e., these are appended to the selected PC scores representing the microstructure statistics as additional features).

Local Gaussian Process Surrogate Models for OPVs

The microstructure PC scores as well as the two distance-based features are used as inputs to train a local Gaussian process (laGP) surrogate model to predict the short circuit current of OPV microstructures. Each input is scaled to exhibit the same variance across the entire ensemble of the dataset. This is needed because laGP models identify local subsets of the training data using suitable distance measures. For each target point, the first n_0 closest neighboring points are chosen as the initial training set for building the initial GP. Subsequently, the ALC criterion is used to sequentially update the training data to maximize the expected information gain. As the training subset is sequentially updated, one expects to see a systematic decrease in the improvement to the model performance. Consequently, one would naturally reach a point where further updating the training set would only minimally improve the laGP model performance. In this study, the sequential update of the laGP model was continued until the reduction in the prediction variance was smaller than 10^{-6}. In the protocol described above, the final size of the local training set is denoted as $n_d = n_0 + n_{ALC}$, where n_{ALC} denotes the number of training points selected using the ALC criterion. The performance of the trained laGP models produced in this work was quantified using multiple error measures, including normalized mean absolute error (*nMAE*), normalized median absolute deviation (*nMAD*) and R^2. These are defined as

$$nMAE = \frac{\frac{1}{N}\sum_{i=1}^{N}\left|J_{sc}^{(i)} - \tilde{J}_{sc}^{(i)}\right|}{\bar{J}_{sc}} \tag{8}$$

$$nMAD = \frac{median\left(\left|J_{sc}^{(1)} - \tilde{J}_{sc}^{(1)}\right|, \left|J_{sc}^{(2)} - \tilde{J}_{sc}^{(2)}\right|, \ldots, \left|J_{sc}^{(N)} - \tilde{J}_{sc}^{(N)}\right|\right)}{\bar{J}_{sc}} \tag{9}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(J_{sc}^{(i)} - \tilde{J}_{sc}^{(i)}\right)^2}{\sum_{i=1}^{N}\left(J_{sc}^{(i)} - \bar{J}_{sc}\right)^2} \tag{10}$$

where $J_{sc}^{(i)}$ and $\tilde{J}_{sc}^{(i)}$ are the actual (ground truth) and the predicted short circuit current of the i^{th} target point, and N is the number of test points. $\bar{J}_{sc}$ denotes the mean value of the J_{sc} values. R^2 serves as an indicator of how much of the variation in the output is explained by the inputs. The value of R^2 for a perfect model is expected to be one. Likewise, for a flat line model that always predicts the mean, the value of R^2 will be zero.

Results and Discussion

In the present study, an ensemble of 33,552 distinct OPV microstructures was generated to establish the desired data-driven microstructure–property linkage for OPV films. The short circuit current J_{sc} associated with each microstructure was obtained by solving the XXD equations discussed in "Microstructure and Photovoltaic Property Dataset" section. The unsupervised feature engineering framework described in "Material states in OPV microstructures" section was employed on each microstructure.

Figure 3 depicts the eight sets of spatial auto-correlations computed for the example microstructure shown in Fig. 2a. The top and bottom rows in this figure present spatial auto-correlations of the four main material states and the four interface states, respectively. Note that the auto-correlations exhibit centro-symmetry, because the values of the statistics for $\boldsymbol{r}$ and $-\boldsymbol{r}$ are the same. Therefore, half the information in these maps is redundant and could be eliminated before performing the PCA. The central peak value in each auto-correlation map, corresponding to $\boldsymbol{r} = \boldsymbol{0}$, reflects the volume fraction of the specific material state. For the interface states, this value corresponds to the volume fraction occupied by the 2-voxel wide interface regions defined in this work. The auto-correlation maps implicitly capture a significant amount of statistical information on the shape, size, and spacing distributions of the material states in the microstructure. For instance, the bands in the $f_{\boldsymbol{r}}^{D^{\Lambda}D^{\Lambda}}$ map capture important features related to the size, shape, orientation, and spacing of the D^{Λ} regions in the microstructure (compare the auto-correlation map with the actual microstructure in Fig. 2a). Similarly, $f_{\boldsymbol{r}}^{A^{\circ}A^{\circ}}$ captures the details of the more compact and isolated positioning of the A° regions in this microstructure. In contrast, the auto-correlation maps for D° and $A^{\vee}$ indicate that these regions are more broadly distributed in the microstructure. Similar observations can be made for the auto-correlation maps of the interface states.

In order to efficiently compute the PCA of the large data matrix of size $33,552 \times 81,608$ assembled in this work, we took advantage of the randomized SVD algorithm implemented in DASK package in Python programming language [43]. It was decided to truncate the PC representations obtained from this protocol to 10 PCs, because there was no appreciable improvement in the variance

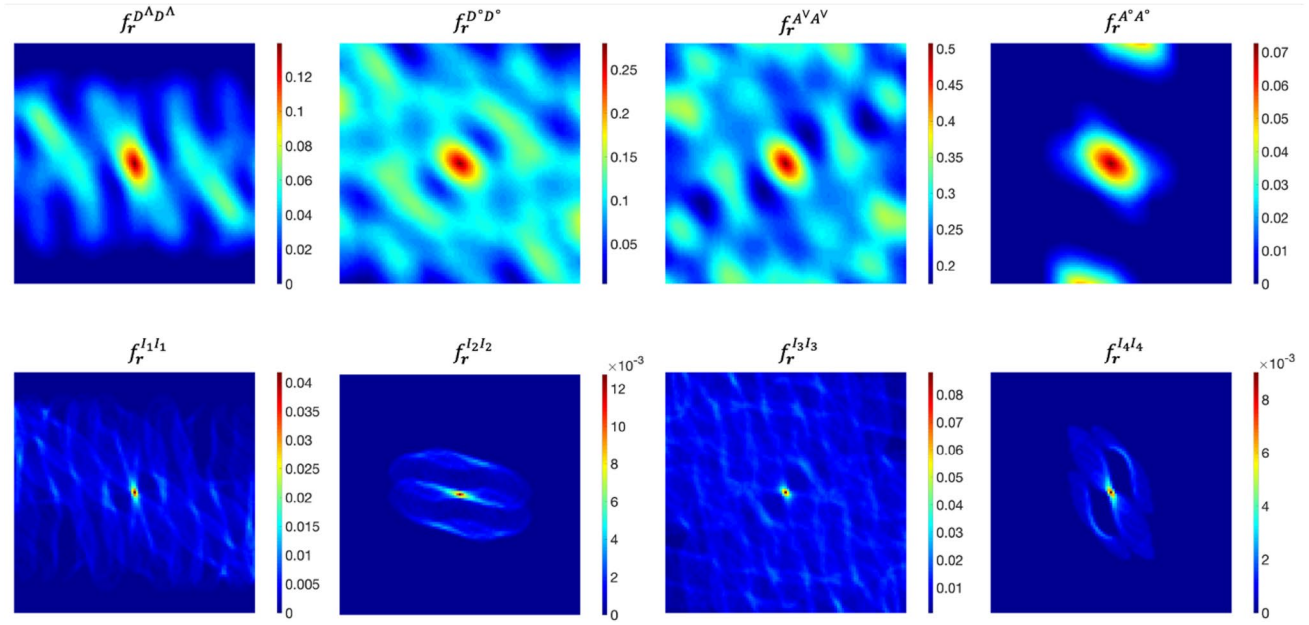

Fig. 3 The 8 sets of 2-point spatial auto-correlations corresponding to main material states and interfaces of the example microstructure shown in Fig. 2 is shown. The center value of these statistical maps is volume fraction of the corresponding material state

captured beyond this truncation level. This represents a significant reduction in the dimensionality of the microstructure representation, where we started with 81,608 spatial correlations and ended up with only 10 PC scores. The representation of all 33,552 microstructures in the first three PCs is presented in Fig. 4. In this figure, each data point corresponds to the first three PC scores of the microstructure statistics and is colored using its value of J_{sc}. Although the three PC scores represent only a subset of the regressors we intend to use in this work (a total of ten PC scores and two distance-based metrics will be used), it is very encouraging to see the patterns in Fig. 4 suggesting a strong dependence of the target on these regressors.

A direct interpretation of the PC scores is currently not possible. Essentially, each PC basis represents a linearly weighted collection of 81,608 spatial correlations. The PC score of each OPV microstructure represents the projection (i.e., dot product) of its set of 81,608 spatial correlations on the corresponding PC basis. The high dimensionality of the PC basis makes it impractical to seek the precise physical meaning of the PC scores. However, it was found that the first PC score is highly correlated to the volume fractions of the four main material local states

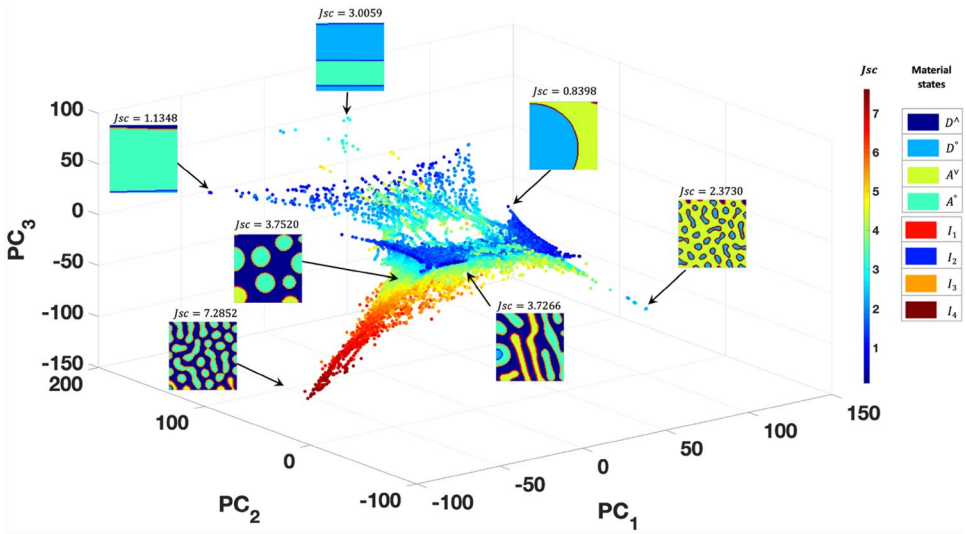

Fig. 4 The low-dimensional representation of the entire data ensemble of OPV microstructures in the first 3 PC basis is depicted. The PC representations are truncated after the first 10 PCs. The unsupervised PCA is powerful in capturing the microstructural differences as well as the variance in the values of J_{sc}

as well as the I_1 and I_3 interface states (interfaces of D^{Λ} with acceptor material states). The second PC score was found to be highly correlated to volume fractions of I_2 and I_4 interface states (interfaces of $A^{\vee}$ with donor material states). In addition to the information on the volume fractions, PC scores contain rich information on other morphological aspects of microstructures such as shape, size, orientation and spacing of material features within OPV microstructures. For instance, as shown in Fig. 4, the microstructures comprising coarser regions of $A^{\vee}$ and/or D° have higher PC_1 values, while the microstructures with coarser regions of A° and/or D^{Λ} have smaller PC_1 values. Several other similar qualitative observations can be made by inspecting Fig. 4 closely. As another example, it can be seen that microstructures comprised mainly/only from coarser unconnected donor/acceptor regions are completely separated from the microstructures with connected donor/acceptor finer regions in the low-dimensional PC representation.

The 10 microstructure PC scores and the two averaged distance-based metrics, δ^{Λ} and $\delta^{\vee}$, are used as inputs to train the surrogate laGP models using the R package language [38]. As already noted, each input feature is scaled to exhibit the same variance across the entire dataset for this model building strategy. A laGP model is produced for each test point using a set of $n_0 = 35$ closest neighbors in the input domain. The ALC criterion is employed to sequentially add points to the design space such that their addition maximizes the expected information gain. The training size for the 33,552 laGP models produced in this study was in the range $[36, 346]$. The distribution of the training sizes is shown in Fig. 5.a. It is seen that more than 99% of the laGP models built in this study needed less than 100 local training data points. This small size of the local training data set significantly reduces the computational cost involved in building the desired laGP models. Figures 5b and 5c present the parity plot comparing the J_{sc} predictions from the laGP models with their corresponding ground-truth values as well as the uncertainty associated with the model predictions (i.e., one standard deviation from the mean prediction shown as error bars) and the distribution of the relative mean absolute errors, respectively. The standard deviation in 99.7% of the trained models is within 5% of the $\overline{J}_{sc}$. Those few models that exhibit higher uncertainties correspond to the microstructures that fall on the boundary of the input PC domain. This is to be expected as laGP performs better in the interior of the input domain, compared to the edges of the input domain (there is limited availability of training points in these regions). The normalized absolute error was higher than 0.05 in only 8% of the trained laGP models. Considering the entire set of trained models, the

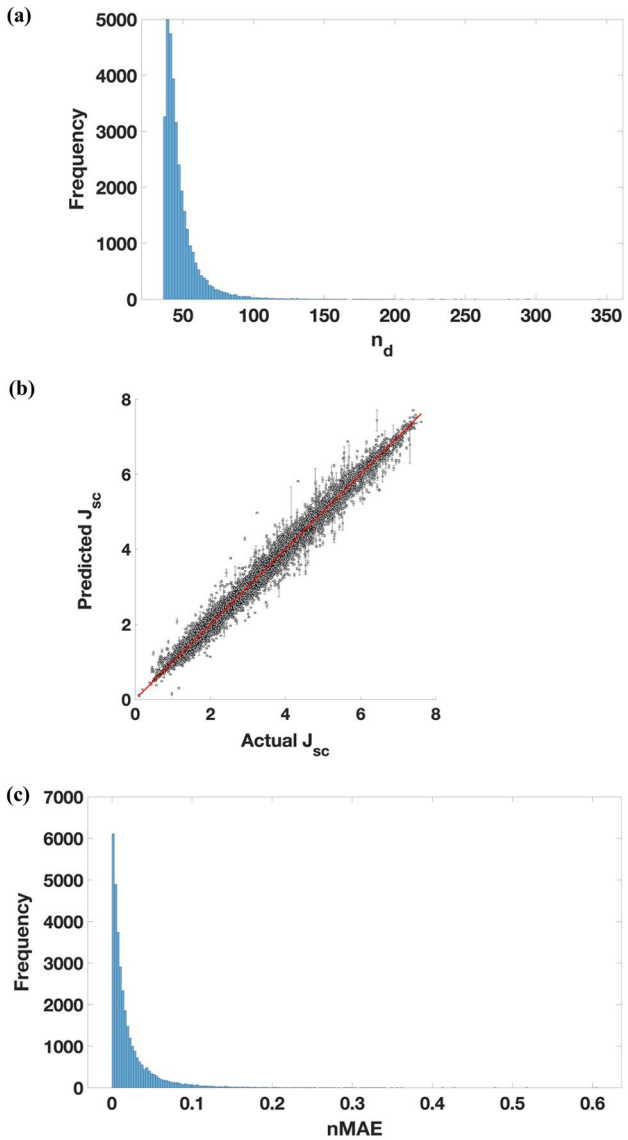

Fig. 5 Depiction of the performance of the data-driven structure–property linkages trained for OPV microstructures is presented. The total design size of each laGP model is determined using ALC criterion. The distribution of the final design size is shown in **a**. The parity plot comparing the predictions and actual values of J_{sc} as well as the relative mean absolute error are presented in **b** and **c**, respectively. The established high-fidelity microstructure–property linkages demonstrate the utility of the developed feature engineering framework for organic photovoltaics

normalized mean absolute error *nMAE* and the normalized mean median absolute deviation *nMAD* were 2.16% and 1.10%, respectively. Moreover, a high value of $R^2 = 0.99$ was calculated for the trained laGP models, which demonstrates that a high proportion of the variance in the target is being captured well by the model inputs. This clearly demonstrates the efficacy of the novel feature engineering framework presented in this work in establishing

high-fidelity data-driven microstructure–property mappings in organic photovoltaics.

Conclusions

A novel unsupervised feature engineering framework for data-driven mappings of OPV microstructures to their functional properties has been successfully developed. This new feature engineering framework successfully leveraged digital image processing algorithms in conjunction with the previously established MKS framework. Specifically, a computationally efficient labeling of two sets of salient material states (four bulk material states and four interface states) was found to be the critical first step in the feature engineering of the OPV microstructures. One set of features characterized the connectivity of the bulk phases controlling the generation and transport of excitons, while the other set of features characterized the interfaces between the bulk phases controlling the generation and transport of charges. This was then followed by the computation of suitable 2-point spatial auto-correlations using the MKS framework, and their low-dimensional representation by PCA performed using a scalable randomized SVD algorithm. In addition to material PC scores, it was found that two additional expert-defined distance-based metrics were essential to improve the accuracy of the data driven structure–property linkages for microstructures where the donor/acceptor pixels are not in direct contact with their corresponding electrodes. Finally, a localized-version of the Gaussian process (laGP) was employed to extract the desired reduced-order structure–property linkages. It was found that the implementation of laGP with ALC produced many computational benefits for the present application. It was shown that with only a small subset of the training dataset one can build accurate laGP models at low computational cost. The uncertainty associated with the model predictions was quantified by considering one standard deviation from the mean prediction. It was found that only 0.3% of the model predictions exhibited a standard deviation higher than 5% of the mean value of the target property. The high-fidelity structure–property linkages extracted in this study attest to the tremendous efficacy of the proposed novel feature engineering framework for complex organic photovoltaic microstructures.

Acknowledgements The authors acknowledge funding from NASA Langley Research Center. BG acknowledges partial support from NSF 1906194 and ONR Award N00014-19-1-2453. SK acknowledges partial support from Vannevar Bush Fellowship through ONR Award N00014-18-1-2879.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Wodo O et al (2013) Quantifying organic solar cell morphology: a computational study of three-dimensional maps. Energy Environ Sci 6(10):3060–3070
2. Wodo O et al (2012) Computational characterization of bulk heterojunction nanomorphology. J Appl Phys 112(6):064316
3. Wodo O et al (2015) Automated, high throughput exploration of process–structure–property relationships using the mapreduce paradigm. Materials Discovery 1:21–28
4. Pokuri BSS., et al., (2019) Interpretable deep learning for guided microstructure-property explorations in photovoltaics. npj Comput Mater 5(1):1-11.
5. Kalidindi SR (2015) Hierarchical materials informatics: novel analytics for materials data. Elsevier.
6. Iskakov A et al (2018) Application of spherical indentation and the materials knowledge system framework to establishing microstructure-yield strength linkages from carbon steel scoops excised from high-temperature exposed components. Acta Mater 144:758–767
7. Latypov MI, Toth LS, Kalidindi SR (2019) Materials knowledge system for nonlinear composites. Comput Methods Appl Mech Eng 346:180–196
8. Yabansu YC et al (2020) A digital workflow for learning the reduced-order structure-property linkages for permeability of porous membranes. Acta Mater 195:668–680
9. Hashemi S, Kalidindi SR (2021) A machine learning framework for the temporal evolution of microstructure during static recrystallization of polycrystalline materials simulated by cellular automaton. Comput Mater Sci 188:110132
10. Torquato S, Haslach H Jr (2002) Random heterogeneous materials: microstructure and macroscopic properties. Appl Mech Rev 55(4):B62–B63
11. Niezgoda SR, Yabansu YC, Kalidindi SR (2011) Understanding and visualizing microstructure and microstructure variance as a stochastic process. Acta Mater 59(16):6387–6400
12. Niezgoda SR, Kanjarla AK, Kalidindi SR (2013) Novel microstructure quantification framework for databasing, visualization, and analysis of microstructure data. Integr Mater Manuf Innov 2(1):54–80
13. Hastie T et al (2005) The elements of statistical learning: data mining, inference and prediction. The Math Intelligencer 27(2):83–85
14. Bishop CM, Nasrabadi NM (2006) Pattern recognition and machine learning. Vol. 4. Springer.
15. Fernandez-Zelaia P, Yabansu YC, Kalidindi SR (2019) A comparative study of the efficacy of local/global and parametric/non-parametric machine learning methods for establishing structure–property linkages in high-contrast 3D elastic composites. Integr Mater Manuf Innov 8(2):67–81
16. Tallman AE et al (2019) Gaussian-process-driven adaptive sampling for reduced-order modeling of texture effects in polycrystalline alpha-Ti. JOM 71(8):2646–2656
17. Yabansu YC et al (2019) Application of Gaussian process regression models for capturing the evolution of microstructure statistics in aging of nickel-based superalloys. Acta Mater 178:45–58
18. Yabansu YC et al (2019) Application of Gaussian process autoregressive models for capturing the time evolution of microstructure

statistics from phase-field simulations for sintering of polycrystalline ceramics. Modell Simul Mater Sci Eng 27(8):084006
19. Parvinian S et al (2020) High-throughput exploration of the process space in 18% Ni (350) maraging steels via spherical indentation stress–strain protocols and Gaussian process models. Integr Mater Manuf Innov 9(3):199–212
20. Marshall A, Kalidindi SR (2021) Autonomous development of a machine-learning model for the plastic response of two-phase composites from micromechanical finite element models. JOM 73(7):2085–2095
21. Williams CK, Rasmussen CE (2006) Gaussian processes for machine learning, vol 2. MIT Press, Cambridge
22. Wodo O et al (2012) A graph-based formulation for computational characterization of bulk heterojunction morphology. Org Electron 13(6):1105–1113
23. Cahn JW Hilliard JE (1958) Free energy of a nonuniform system. I. Interfacial free energy. J Chem Phys 28(2):258-267.
24. Wodo O, Ganapathysubramanian B (2012) Modeling morphology evolution during solvent-based fabrication of organic solar cells. Comput Mater Sci 55:113–126
25. Wodo O, Ganapathysubramanian B (2014) How do evaporating thin films evolve? Unravelling phase-separation mechanisms during solvent-based fabrication of polymer blends. Appl Phys Lett 105(15):153104
26. Zhao K et al (2016) Vertical phase separation in small molecule: polymer blend organic thin film transistors can be dynamically controlled. Adv Func Mater 26(11):1737–1746
27. Kodali HK, Ganapathysubramanian B (2012) Computer simulation of heterogeneous polymer photovoltaic devices. Modell Simul Mater Sci Eng 20(3):035015
28. Kodali HK, Ganapathysubramanian B (2012) A computational framework to investigate charge transport in heterogeneous organic photovoltaic devices. Comput Methods Appl Mech Eng 247:113–129
29. Pfeifer S et al (2018) Process optimization for microstructure-dependent properties in thin film organic electronics. Materials Discovery 11:6–13
30. Fullwood DT et al (2010) Microstructure sensitive design for performance optimization. Prog Mater Sci 55(6):477–562
31. Cecen A, Fast T, Kalidindi SR (2016) Versatile algorithms for the computation of 2-point spatial correlations in quantifying material structure. Integr Mater Manuf Innov 5(1):1–15
32. Khosravani A, Cecen A, Kalidindi SR (2017) Development of high throughput assays for establishing process-structure-property linkages in multiphase polycrystalline metals: Application to dual-phase steels. Acta Mater 123:55–69
33. Paulson NH et al (2017) Reduced-order structure-property linkages for polycrystalline microstructures based on 2-point statistics. Acta Mater 129:428–438
34. Wilson AH, Nickisch R (2015) Kernel interpolation for scalable structured Gaussian processes (KISS-GP). In: Proceedings of the 32nd international conference on machine learning, 1775-1784
35. Bui TD, Turner RE (2014) Tree-structured Gaussian process approximations. Advances in Neural Information Processing Systems. 27
36. Lee B-J, Lee J, Kim K-E (2017) Hierarchically-partitioned Gaussian process approximation. In: Proceedings of the 20th international conference on artificial intelligence and statistics, PMLR, vol 54, pp 822-831
37. Gramacy RB, Apley DW (2015) Local Gaussian process approximation for large computer experiments. J Comput Graph Stat 24(2):561–578
38. Gramacy RB (2016) laGP: large-scale spatial modeling via local approximate Gaussian processes in R. J Stat Softw 72:1–46
39. Heaton MJ et al (2019) A case study competition among methods for analyzing large spatial data. J Agric Biol Environ Stat 24(3):398–425
40. Cohn DA (1996) Neural network exploration using optimal experiment design. Neural Netw 9(6):1071–1083
41. Haralock RM, Shapiro LG Computer and robot vision. 1991: Addison-Wesley Longman Publishing Co., Inc.
42. Shaw PE, Ruseckas A, Samuel ID (2008) Exciton diffusion measurements in poly (3-hexylthiophene). Adv Mater 20(18):3516–3520
43. Rocklin M, Dask (2015): Parallel computation with blocked algorithms and task scheduling. In Proceedings of the 14th python in science conference. Citeseer.