MRS Communications

https://doi.org/10.1557/s43579-021-00147-4

MRS Communications

Research Letter

How important is microstructural feature selection for data-driven structure-property mapping?

Hao Liu, Materials Design and Innovation Department, University at Buffalo, Buffalo, NY, USA
Berkay Yucel, George W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA, USA
Daniel Wheeler, Material Measurement Laboratory, Materials Science and Engineering Division, National Institute of Standards and Technology, Gaithersburg, MD, USA
Baskar Ganapathysubramanian, Mechanical Engineering Department, Iowa State University, Ames, IA, USA
Surya R. Kalidindi, George W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA, USA; School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA, USA
Olga Wodo, Materials Design and Innovation Department, University at Buffalo, Buffalo, NY, USA

Address all correspondence to Olga Wodo at **olgawodo@buffalo.edu**

(Received 2 July 2021; accepted 8 December 2021)

Abstract

Data-driven approaches now allow for systematic mapping of microstructure to properties. In particular, we now have diverse approaches to "featurize" microstructures, creating a large pool of machine-readable descriptors for subsequent structure-property analysis. We explore three questions in this work: (a) Can a small subset of features be selected to train a good structure-property predictive model? (b) Is this subset agnostic to the choice of feature selection algorithm? And (c) can the addition of expert-identified features improve model performance? Using a canonical dataset, we answer in the affirmative for all three questions.

Introduction

The holy grail of materials science is to discover the physically meaningful microstructural features controlling the material properties of interest and to describe such relationships in forms useful for optimal design of engineered components. Therefore, the core materials knowledge is often expressed as structure-property (SP) relationships: $P = F(\widehat{d})$, where P is the property of interest, and $\widehat{d}$ is the vector of salient microstructural features or descriptors. The function F is typically learned via hypothesis-driven experiments or physics-based numerical simulations or, more recently, via machine-learning approaches.[1–4]

Mapping microstructure-sensitive properties with microstructure representation is invariably challenging due to the mismatch between the high dimensionality of microstructural information (e.g., via microscopy or simulations) and the principal degrees of freedom (or salient features) governing the SP models. This is because microstructural imaging aims to provide detailed, high-resolution maps. Hence, imaging techniques inevitably produce high-dimensional representations of microstructure, while the goal of establishing practically useful SP models is to identify the smallest set of features that can successfully predict the effective properties exhibited by the material. Often, this set is not known a priori, especially for complex multi-physics phenomena governing the material properties. Thus, the efficient learning of F depends critically on the availability of a large pool of computable features[1] and principled approaches for selecting the most informative (or salient) features.

There exist several distinct approaches to "featurize" the microstructure. These include voxel-based representations,[3,4] characterization via physical descriptors,[5,6] statistical functions,[7,8] spectral density functions (SDF),[9,10] and machine-learning methods.[1,2] Features may include physically meaningful descriptors (e.g., grain size, volume fraction, tortuosity), statistical function (e.g., two-point correlation), or local neighborhoods. Regardless of the features, the microstructure (typically an image in 2D or image stack in 3D) is converted into a machine-friendly format one can subsequently perform computations upon. For a detailed comparative discussion of various representations, we refer the reader to recent review papers.[5,11]

Our motivation here is to understand the importance of data representation and subsequent selection of salient features, as well as the robustness of unsupervised selection of salient features. Additionally, we evaluate the utility of including selected expert-enriched features (i.e., domain knowledge) in enhancing the predictive power of the trained models.

We utilize a problem of constructing SP models for organic photovoltaics applications (OPV). It is well known that the microstructure of OPV active layers determines, to a large extent, the photovoltaic performance of the device. Hence, there is a critical need to establish SP models for this application. We

[1] We use the words "features" and "descriptors" interchangeably.

utilize an open-source dataset[12] of microstructures and OPV properties—specifically, short circuit current J_{sc}—as our canonical dataset. We utilize human-derived and machine-derived features with machine-learning approaches to construct SP models. Our benchmark for comparison is a SP model carefully derived using expert-derived features. We explore how well machine-derived features can emulate an expert in deriving the salient features for this specific SP mapping. Finally, the paper is supplemented with a set of notebooks showcasing the basic steps involved in constructing SP models (see section Data availability).

Materials and methods

This work examines methods for constructing SP maps for OPV applications. The focus of modeling is on the effect of microstructure on the OPV device performance. The microstructure constitutes the active layer of OPV. It consists of two phases, where one phase serves as an efficient electron-donor, and the other serves as an efficient electron-acceptor. The active layer being modeled is sandwiched between two electrodes: an anode and a cathode. In this work, the performance of an OPV device is characterized by its short circuit current, J_{sc}. The J_{sc} is derived using a physics-based computational model that solves the excitonic drift-diffusion equations. The model focuses on the charge transport through the microstructure (based on a well-studied material system, P3HT:PCBM blend[2] mixture). The model solves for the spatial distribution of excitons, electrons, holes, and the electric potential across the active layer of the OPV device. The ML models are trained on an open-source dataset with 1708 OPV microstructures generated using a Cahn–Hilliard equation solver.[13] Each microstructure in this dataset is a two-dimensional, two-phase microstructure of size 401×101 pixels and is annotated by one property (J_{sc} from the computational model). Additional details on data generation and the computational models are presented in the supplementary information and our prior work.[13,14]

The dataset is of moderate size, but predicting properties required substantial resources.[15] However, the major reason behind selecting this dataset is the availability of the SP model derived by the expert. In the paper, we refer to it as a reference model M_E. It is also a data-driven model trained previously on the same dataset. However, domain experts first established this model by defining a large set of potential features and then identifying three salient features through repeated trial and error correlation studies. This model is used in this paper to compare with other methods of feature selection. We note that model M_E should not be considered as a ground-truth model but rather a reference model as the name suggests.

The short circuit current is our property of interest and the ground-truth values for J_{sc} are computed using the computational model, and then used to determine the accuracy of the machine-learning (ML) models examined in this paper. All models are trained, tested and validated with a data split into a training set (80%) and a testing set (20%). The performance of each model is evaluated by fivefold cross validation (performed on the training set).

We compare the performance of SP models built with different data-driven featurization schemes against model M_E. The accuracy of all the ML models examined in this paper are evaluated by comparing against the ground-truth data (which is computed from a detailed physics-based model). All the ML models in this paper can be thought of as surrogate models to this detailed physics-based model.[14]

Three levels of microstructure representations

Formally, we consider three microstructure data representations levels (RL): the raw data (RL0), the featurized data (RL1), and the extracted salient features (RL2)—see Fig. 1.

Representation layer zero (RL0): The raw data (i.e., image data) constitute the RL0. The raw data size depends on the resolution and size of the sample. While it is possible to train SP models that directly map raw data to output,[1,16] several challenges exist—including the curse of dimensionality[17] that necessitates the availability of very large datasets and the "black box" nature of such models, which makes extracting scientific insight non-trivial. Additionally, it is non-trivial to enforce underlying invariances (e.g., translation and/or rotation invariance) that could play a part in determining the output. An extra layer of representation is introduced (RL1) to overcome these challenges. We refer to this step as the featurization step (RL1). We formally denote the raw dataset of N microstructures as $\mathcal{X} = \{X_1, \ldots, X_N\}$, where microstructure X_i is represented by a $(n_x \times n_y)$ bitmap with bitmap pixel $X_i(x, y) \in \{0, 1\}$ at position (x, y).

Representation layer one (RL1): This level corresponds to the feature layer, where transformations are applied to RL0. Here, we consider two classes of features: human-derived and machine-derived features (see Fig. 1). The human-derived features consist of application-specific descriptors.[18] Such descriptors require input from experts to formulate and compute. While this featurization approach carries the risk of missing key features due to unintended bias or lack of information, the feature set is typically physically meaningful, explainable, and interpretable. Examples include volume fractions, interfacial area per unit volume, connected components density, average domain sizes, tortuosity of the paths, and percent contact area with boundaries. The dimensionality of this feature set is usually much smaller than the dimensionality of the input microstructure. In this work, we use twenty-one descriptors computed using a graph-based approach[18] to form two vectors of descriptors. The first vector d consists of nineteen descriptors defined based on an understanding of photophysics operations. The second vector d' is appended with two additional descriptors enriched by the expert. We refer to these two vectors as $d = \{d_1, \ldots, d_l : d_i \in \mathbb{R}\}$ with

[2] P3HT:PCBM is poly(3-hexylthiophene) and 1-(3-methoxycarbonyl)-propyl-1-phenyl-[6,6]C_{61}.

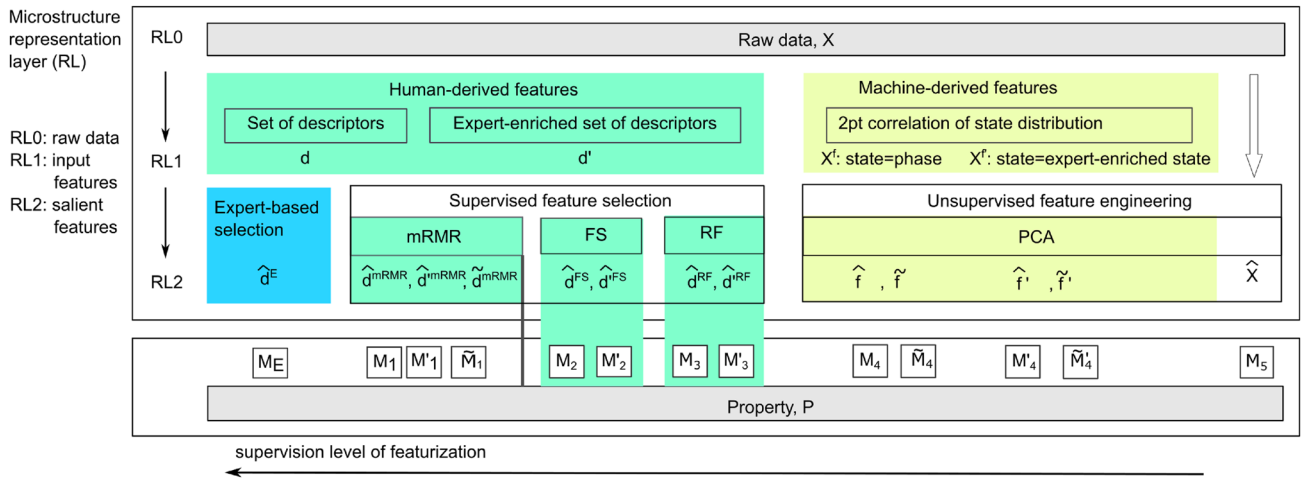

Figure 1. A taxonomy of microstructure representations with three layers RL0, RL1, RL2 allows principled classification of various approaches to construct SP models. RL0 consists of raw data that can be featurized into RL1 using two approaches: human-derived features (here descriptors) and machine-derived features (here two-points spatial correlations). Both types of features can be enriched with expert knowledge. In the last feature engineering layer—RL2, the salient features are extracted using three types of approaches: expert-based selection (blue box), supervised feature selection (green boxes), and unsupervised feature engineering (yellow boxes). The salient features from RL2 are used to construct models of varying complexity. Models are labeled M_E, M_1 to M_5 as shown schematically. Note that all models with a tilde use the monomial-augmented salient features, and all models with prime superscripts correspond to the expert-enriched features.

cardinality of $l = 19$ and the expert-enriched features d' with cardinality $(l + 2)$. Specifically, three stages of photocurrent generation—light absorption, exciton dissociation, and charge transport—guide the definition of these descriptors. We refer to [19] for a detailed description of these descriptors and Supplementary Information for the list of descriptors. The two additional expert-defined descriptors were based on an in-depth, time-consuming sensitivity and correlation analysis of the full-physics simulations that predict the short circuit current. We emphasize that these descriptors (contact area of donor phase with anode and acceptor phase with cathode) are non-trivial, expert knowledge-enriched descriptors. Our descriptors range from fairly straightforward, like volume fraction, to descriptors defined by the experts, like the last two descriptors mentioned above. A general rule of machine learning is that starting with more descriptors often results in an improved model, as tools for feature selection can screen irrelevant descriptors and then construct a model using only the down-selected features. However, the number of features should be sufficient to capture the underlying relationship between descriptors and property with good accuracy and without overfitting.

For the machine-derived features, two-point spatial auto-correlations (also known as two-point statistics) are generated using the open-source package, PyMKS (The Materials Knowledge System in Python), see [20,21]. There is an extensive literature on using spatial distributions to represent microstructures for structure-property models.[22–24] For the two-phase material system under consideration, only one auto-correlation of the electron-accepting phase is needed.[25,26] Consider a microstructure, X_i. Let m_s denote this microstructure as an array, where s indexes each pixel, and the values of m_s reflect the volume fraction of the electron-acceptor phase in the pixel s. In the microstructures considered in this work, each pixel is fully occupied on one of the two phases present in the microstructure. Hence m_s takes values of zero or one. The auto-correlation of interest is defined as:

$$f_r = \frac{1}{S_r}\sum_s m_s m_{s+r} \text{ and } F_i = \{f_r \,\forall r \in S_r\} \qquad (1)$$

where f_r denotes the auto-correlation array indexed by a set of discrete vectors r. S_r represents the total number of valid placements of the discrete vector r used in evaluating the spatial statistics,[22,27] and F_i corresponds to auto-correlation array of microstructure X_i in $\mathcal{X}$. The two-point correlation is computed for each microstructure, X_i in $\mathcal{X}$ and then aggregated to form the machine-featurized dataset $X^f = \{F_1, F_2, ..., F_N\}$.

Machine-derived features can also be enriched with expert knowledge by assigning new material states to each pixel. Specifically, the output J_{sc} is known to depend on the availability of donor phase adjacent to the anode and acceptor phase adjacent to the cathode. To account for this, a 1-D auto-correlation of the acceptor phase on the surface adjacent to the cathode and a 1-D auto-correlation of the donor phase on the surface adjacent to the anode are added to the earlier feature set, f, to generate the enriched feature set, f'. Mathematically, $f' = \{f, f_{An}, f_{Ca}\}$, where f_{An} and f_{Ca} are appropriate 1-D auto-correlations on the layers adjacent to the anode and cathode, respectively. Similar to machine-featurized dataset, the expert-enriched dataset is formed $X^{f'} = \{F'_1, F'_2, ..., F'_N\}$, where F'_i corresponds to

auto-correlation array of expert-enriched state in X_i microstructure.[3] It is important to note that the two-point statistics reflect the directional dependencies of the extracted microstructure measures. However, the dimensionality of this feature array is large—same order as the input microstructure.

Representation layer one (RL2): This layer corresponds to a "concentration of information", where the number of features is reduced, ideally without degrading the predictive power of the SP model being built. This is an important step in constructing surrogate models, because interpolation theory suggests that for a fixed number of samples, more accurate interpolants can be constructed when the number of features is smaller.[28,29] However, identifying a proper salient features is challenging. First, there is no uniqueness guarantee for a set of salient features. Different data-driven approaches could result in different sets of salient features. Second, the set of salient features can be incomplete.[4] Finally, there is no guarantee that the salient features are interpretable, thus precluding an easy generation of insight.

We broadly identify three approaches used for salient feature selection: (a) expert-based selection, (b) supervised feature selection, and (c) unsupervised feature engineering. Figure 1 visually lays out this classification at RL2. The first two approaches are used on human-derived features, while the last approach is applied on machine-derived features.

In the first approach, an expert defines the vector of salient features, denoted as $\widehat{d}^E$. In Fig. 1, this approach is marked with the blue box. The vector $\widehat{d}^E$ ($\widehat{d}^E \in d'$) can be derived using a hypothesis-driven approach or, as in our case, an unsupervised approach relying on the correlation studies. Here, the vector of expert-derived salient features consists of three features: $\widehat{d}^E = \{d_{10}, d_2, \min(d_{20}, d_{21})\}$. These salient features are d_{10}—the volume fraction of electron-donor phase (as this is the phase that contributes to the light absorption), d_2—the weighted fraction of the electron-donor phase (where weighting is applied to the shortest distance to the interface and captures the efficacy of exciton diffusion), and finally $\min(d_{20}, d_{21})$ the minimal contact area with the electrode (donor with an anode, and acceptor with cathode). The product of three descriptors correlates well with the short circuit current, but identifying this vector of features required tedious, manual, and time-consuming investigations by a domain expert. These three features are used to construct our reference model M_E (cf. Fig. 1).

In the second approach, three types of off-the-shelf feature selection techniques[30] are applied: filter methods, wrapper methods, and embedded methods. For each type of method, we choose one technique that we briefly describe here in the main document and provide more details in Supplementary Information. The filter methods are the simplest to use. Here, we select the maximum Relevance Minimal Redundancy method (mRMR). Iteratively, this technique seeks to select down a small set of features that have a strong correlation with the targeted properties and a low redundancy with other features selected in previous iterations. It is a relatively simple technique that does not involve any SP model construction but only looks at the basic correlation between variables (either descriptors/features or property). As a result, the input features are ordered based on their score, capturing relevance and redundancy. The scoring is then used to decide on the number of salient features used as inputs to the SP model. Once the number of salient features is selected, any model construction strategy can be used.

In wrapper methods, feature selection and machine learning are coupled. Forward selection (FS) is a representative method used in this paper, and it involves an iterative process that starts with an empty set of features. In each iteration, the feature that best improves the model is added until the addition of a new feature does not improve the model's performance. Because these two steps are linked, this type of method is prone to overfitting. Moreover, for a large pool of descriptors, wrapper methods may be computationally demanding.

Finally, embedded methods, such as Random Forrest (RF) used in this work, are the most complex feature selection methods. The Random Forest algorithm constructs hundreds of decision trees, each building a SP mapping over a random extraction of the observations from the dataset with a random combination of the input features. As a consequence, RF is less prone to overfitting at the expense of computational cost. The feature selection (or scoring) is computed from each decision tree and averaged over all the trees. When training an individual tree, the output variance is minimized at each node of the particular tree. In this method, an individual feature's relevance is based on the decrease in its variance.

All feature selection methods described above are characterized as supervised methods. In some cases, the salient feature selection is tightly coupled with the model. This is the case for forward selection and random forest models. In Fig. 1, we highlight the tight link by drawing the green boxes around salient features and the models. In filter methods, the model construction is independent of the feature selection. The importance score is computed to capture correlation with the property. In Fig. 1, we highlight the weak link between model and property with a vertical line. In this paper, mRMR, FS and RF are applied to two vectors of descriptors d and d' to derive the corresponding salient feature vector $\widehat{d} = \{\widehat{d}_1, \widehat{d}_2, ..., \widehat{d}_S\}$ of size $S \ll l$. The vectors with salient features are denoted with the superscript of the method used: $\widehat{d}^{\mathrm{mRMR}}$ and $\widehat{d}^{\prime\mathrm{mRMR}}$, $\widehat{d}^{\mathrm{FS}}$, $\widehat{d}^{\prime\mathrm{FS}}$, $\widehat{d}^{\mathrm{RF}}$, and $\widehat{d}^{\prime\mathrm{RF}}$. The vectors of salient features with superscript prime are selected from the vector $\widehat{d}'$ (e.g., $\widehat{d}^{\prime\mathrm{mRMR}} = \{\widehat{d}_1, \widehat{d}_2, ...\widehat{d}_S : \widehat{d}_i \in d'\}$), and the vectors of salient feature without the superscript are selected from the vector d (e.g., $\widehat{d}^{\mathrm{mRMR}} \in d$).

In the third approach—unsupervised feature engineering—salient features are identified independent of the properties. Principal Component Analysis (PCA) is commonly used for this purpose. PCA seeks to rotationally transform

[3] Expert-enriched features are scaled through standardization (or Z-score normalization) before the feature engineering step to eliminate bias toward the subset with highest variance.

[4] Completeness is difficult to confirm even for hypothesis-driven selection approaches.

the data while maximizing the variance capture in the dataset in any selected (low) number of dimensions. The number of salient features is chosen based on the unexplained (residual) variance in the dataset (or can be selected using the performance of the SP model). In this paper, we apply PCA to the machine-derived features (X^f and $X^{f'}$) from RL1 to derive salient features $\widehat{f} = \{\mathrm{PC}_1(X^f), ..., \mathrm{PC}_S(X^f)\}$ and $\widehat{f'} = \{\mathrm{PC}_1(X^{f'}), ..., \mathrm{PC}_S(X^{f'})\}$, where $\mathrm{PC}_i(X^f)$ is the principal score of X^f and $\mathrm{PC}_i(X^{f'})$ is the principal score of $X^{f'}$. We also apply PCA directly to the raw data from RL0 (X) to derive $\widehat{x} = \{\mathrm{PC}_1(\mathcal{X}), ..., \mathrm{PC}_S(\mathcal{X})\}$.

Microstructure-property models

This study aims to compare the predictive power of the different salient features computed in RL2. Toward this goal, different SP models are constructed and their predictive accuracy is evaluated (see also Fig. 1). The model M_E is considered as the reference model. The parametric form of this model is grounded in the process of the current generation in OPV that involves a sequence of three steps, where the outcome of each subsequent step depends on the previous step. Hence, the property (P or J_{sc}) is modeled as the product of powers of three salient features: $P = \prod_{i=1}^{3} \widehat{d}_i^{A_i}$. Taking logarithm of both sides, the model form is expressed as $\log(P) = \sum_{i=1}^{3} A_i \log(\widehat{d}_i)$, allowing the application of linear regression methods for building the desired model.

All other models in the paper are also constructed using linear regression. In the simplest case, the model takes the form of: $P = \sum_i^S A_i \widehat{d}_i$, where $\widehat{d} = \{\widehat{d}_1, \widehat{d}_2, ..., \widehat{d}_S\}$ are the salient features, and $A = \{A_0, A_1, ..., A_S\}$ are the influence coefficients capturing the SP map. This form is used in models M_1, M_2, M_3, M_1', M_2', M_3', M_4, M_4', and M_5. In the first six models, the salient features correspond to the human-derived salient features, in the next two models the salient features corresponds to the PC scores of machine-derived features ($\widehat{f}$, $\widehat{f'}$). Model M_5 uses the PC scores obtained directly from the raw data ($\widehat{x}$). We also augment the salient features $\widehat{d}$ with monomials of order up to q. Here, the surrogate model is expressed as $P = \sum_q^{\widetilde{S}} A_q \widetilde{d}^q$, where $\widehat{d}^q = \widehat{d}^{q_1}\widehat{d}^{q_2}...\widehat{d}^{q_{\widetilde{S}}}$ are power products (monomials) of the salient features, and $A = \{A_0, A_1, ..., A_{\widetilde{S}}\}$ are the influence coefficients capturing the SP map. For example, for vector $\widehat{d} = \{\widehat{d}_1, \widehat{d}_2\}$, and $q = 2$, the extended vector of features (monomials) would be: $\widetilde{d} = \{1, \widehat{d}_1, \widehat{d}_2, \widehat{d}_1^2, \widehat{d}_2^2, \widehat{d}_1\widehat{d}_2\}$. The size of the extended feature vector is $\widetilde{S}$. The monomial-augmented vectors are formed for salient features $\widehat{d}^{\mathrm{mRMR}}$, $\widehat{d}'^{\mathrm{mRMR}}$ and $\widehat{f}$, $\widehat{f}'$ to form the corresponding vectors $\widetilde{d}^{\mathrm{mRMR}}$, $\widetilde{d}'^{\mathrm{mRMR}}$ and $\widetilde{f}$, $\widetilde{f}'$. The SP maps constructed using the extended vectors of features are denoted with tilde and include $\widetilde{M}_1$, $\widetilde{M}_1'$, $\widetilde{M}_4$ and $\widetilde{M}_4'$. In the first two of these models, the salient features correspond to the descriptor vectors $\widetilde{d}^{\mathrm{mRMR}}$ and $\widetilde{d}'^{\mathrm{mRMR}}$, while for the last two models, the salient features correspond to the vectors $\widetilde{f}$ and $\widetilde{f}'$.

Figure 1 lays out all models below the corresponding salient features and methods used to derive them. Note that all models with the prime superscript correspond to the expert-enriched features (d', f'), while models without these superscripts correspond to the smaller features (d, f). Models with subscripts one, two, and three use the salient features from mRMR, FS, and RF, respectively. All models with subscript four correspond to the machine-derived features projected into a low dimensional embedding using PCA. Model with subscript five operated directly on the raw data projected into low dimensional embedding using PCA. All models with tilde correspond to the features space augmented with monomials.

Results and analysis

This study compares the capabilities of various featurization methods (creation, selection, engineering) of the microstructure to enable robust data-driven SP mapping between OPV microstructure and its short circuit current. Altogether, thirteen different SP models are constructed and evaluated against each other. Table I presents a summary of the comparisons, while Figs. 2 and 3 depict the results of feature selection and feature engineering methods on the SP model performance. We report the number of salient features S, the salient features, and the performance measure (R^2) for prediction.[5] The accuracy is evaluated using the physics-based computational model as the ground-truth.

Our results indicate that model M_E offers very good performance among all models built. This model is the expert-derived model with only three salient features manually selected by the expert in repeated trials. The R^2 value for model M_E is 0.97. The superior performance is consistent across training, validation, and prediction. Nevertheless, models with comparable performance can be constructed for both human-derived and machine-derived features only if features are enriched with expert knowledge and suitably augmented (using monomials in the present study). Note the good performance of models ($\widetilde{M}_1'$ and $\widetilde{M}_4'$ in Table I). Next, we compare and contrast various settings of model construction to answer the three major questions raised in the abstract.

We begin by comparing various feature selection methods to construct a SP model. Figure 2 depicts the results for models M_1', M_2', M_3' for the expert-enriched vector of descriptors d'. Three panels of the figure depict the order of descriptors from d' with the decreasing importance score (except for the forward selection method where the score is not computed explicitly). Each panel also reports the change in model performance as one includes systematically the identified important features in the model building. Clearly, the improvement in the R^2 values is minimal with the less important features. These characteristics are evident from all three panels. For example, in the right

[5] Results in Supplementary Information additionally report the normalized mean absolute errors (NMAEs) for fivefold validation and prediction.

panel of Fig. 2, the R^2 increases significantly for the first four features, whose importance scores are considerably higher than the rest of the features. These four features correspond to the salient features identified by the method. The selected number of features is marked with the vertical red line for guidance. The number of salient features is chosen based on the gap in the score value between subsequent features (mRMR, RF) or asymptotic behavior of R^2 (FS). Interestingly, these three methods select a similar number of salient features: four or five. Tables 1 and 2 in Supplementary Information provide the number of salient features S for all models. It is seen that all three types of feature selection methods (mRMR, FS, RF) offer a comparable performance of $R^2 = 0.88$–0.93 (see values for models M_1', M_2', M_3' in Table 2 of the Supplementary Information). Moreover, the salient feature/descriptors in the three vectors (see Table I) are consistent. The interfacial area (d_3), the donor contact area with the anode (d_{20}), and the acceptor contact area with the cathode (d_{21}) have been the most commonly selected feature among these three models. These descriptors match or are strongly correlated with the expert-derived salient features (see Fig. 1 in the Supplementary Information). This demonstrates that for this particular application feature selection is agnostic to the selection method. This is important because it demonstrates that when the featurization of microstructure is performed well, the machine-learning model can be constructed with good performance and minimal additional intervention from the domain expert. Moreover, it affirmatively answers the question of whether a small set of features can be selected to train a predictive model of SP.

We note that the features used to construct these models include two expert-crafted features (d_{20} and d_{21}). To ask the question on the importance of expert involvement in the process of creating the features, we constructed the analogous models (M_1, M_2 and M_3) with the smaller pool of descriptors d. For the same number of salient features, the performance decreases consistently across three types of methods (mRMR,

Table I. Performance of SP models using feature selection and feature engineering applied on four types features: expert-derived features (second column), human-derived features (third and forth column), machine-derived features (fifth and sixth column), and raw data (last column).

	Expert-derived features and model	Feature selection (mRMR) on human-derived features		Feature engineering (PCA) on machine-derived features		PCA on raw data
RL1 features	d'	d	d'	f	f'	X
RL2 features	$(d_{10}, d_2, \min(d_{20}, d_{21}))$	$(d_3, d_{10}, d_8, d_{15}, d_{19}, d_{11})$	$(d_3, d_{10}, d_{21}, d_8, d_{20})$	$\widehat{f}$	$\widehat{f}'$	$\widehat{x}$
S ($\widetilde{S}$)	3	6 (28)	5 (21)	7 (36)	7 (36)	11
Model	M_E	M_1 ($\widetilde{M}_1$)	M_1'($\widetilde{M}_1'$)	M_4 ($\widetilde{M}_4$)	M_4' ($\widetilde{M}_4'$)	M_5
Performance	**0.97**	0.81 (0.83)	0.89 (0.98)	0.75 (0.85)	0.87 (0.95)	0.60

Bold value indicates the highest R^2 value among all models.

Model performance is reported as R^2 on testing set.

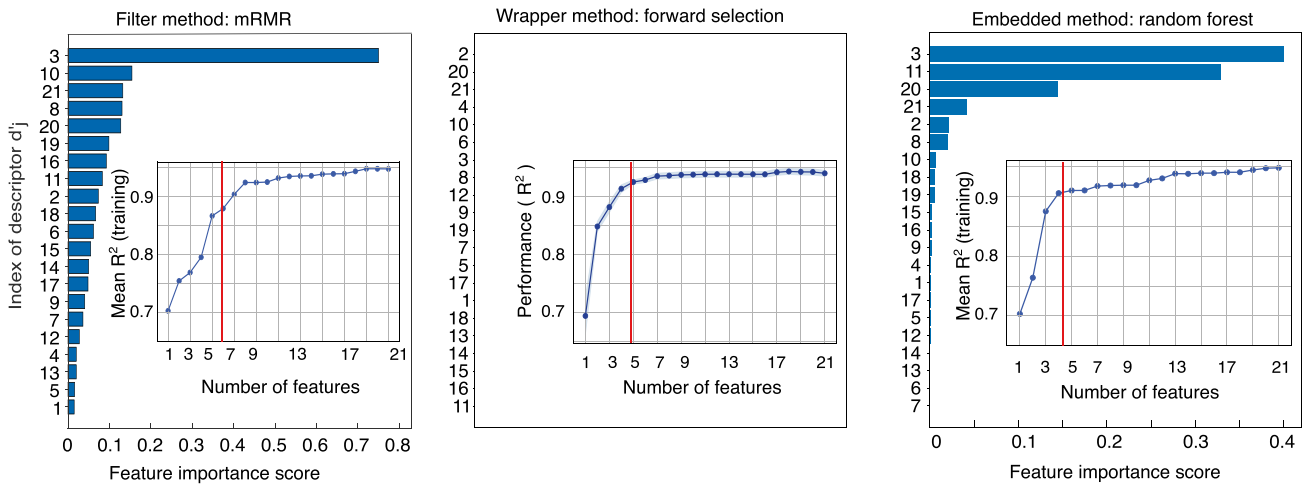

Figure 2. Performance of feature selection on human-derived features d' and the associated model of SP: (left) filter method (mRMR) and model M_1', (middle) wrapper method (forward selection) and model M_2', and (right) embedded method (random forest) and model M_3'. The red vertical line corresponds to the number of features selected for the final model used to report the accuracy values in Table I. Note that all methods consistently choose similar salient features (the number and the descriptors $d_3, d_{20}, d_{21}, d_{10}$) for which models offer optimal performance.

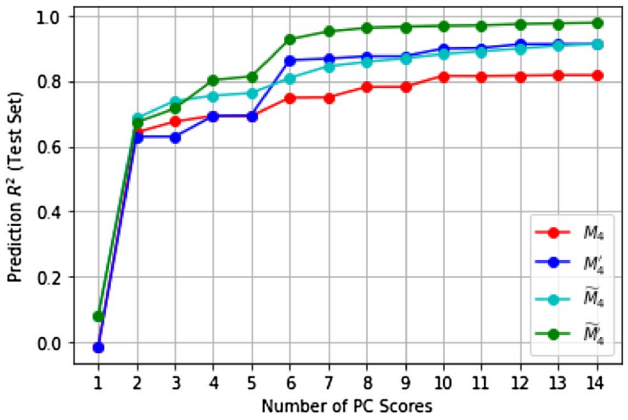

Figure 3. The correlation coefficient against the number of PCA components for different versions of model M_4. The features used in M_4 are derived from two-point correlations of the phase distribution and then transformed via a PCA into a lower dimensional representation (seven components are used for data reported in Table I).

FS, RF). The R^2 of prediction decreased by 0.08 for forward selection, 0.05 for mRMR, and 0.10 for the random forest regression model. Moreover, the performance behavior for an increasing number of features increases only gradually (see Supplementary Information Figs. 4, 6 and 7) without a clear asymptotic behavior. Even with the full set of nineteen features, the built SP model cannot reach a comparable performance of models with salient features selected from the full vector d'. Finally, the lost performance cannot be compensated for by monomial-augmented features. The R^2 values for $\widetilde{M}_1$ model increases by only 0.02 compared to the M_1 model.[6] In contrast, when the same model with augmented features is applied on the feature vector d', the R^2 increases to 0.98 (model $\widetilde{M}_1'$), which is the highest performance among all models. These consistent results reiterate the importance of the input features, and reinforce the value of expert-enrichment to the data representation and featurization. Moreover, our results demonstrate that the addition of expert-derived features can significantly improve the model performance and these improvements cannot be reproduced by increasing the complexity of the model without expert-enriched features.

In the second part of this study, four different SP models (M_4, M_4', $\widetilde{M}_4$, $\widetilde{M}_4'$) with machine-derived featurization are produced and compared. Models M_4 and M_4' are models with regular machine-derived features and expert-enriched machine-derived features. $\widetilde{M}_4$, $\widetilde{M}_4'$ correspond to models with monomial-augmented models with regular machine-derived features and expert-enriched machine-derived features. These SP models are constructed with increasing number of principal components to evaluate their influence on model performance. Figure 3 depicts the performance of these models where the prediction (test set) R^2 scores are presented as functions of the number of PC scores used. The results indicate that the models perform better with increasing number of PC scores (used for model building) up to 7 PC scores. The first seven PC scores are found to explain $> 95\%$ of the entire OPV dataset. Therefore, only the first seven PC scores from each workflow are used for establishing SP models with machine-derived features. As seen from the Table I and Fig. 3, using expert guidance in machine-derived features consistently improved prediction performance on all models. This observation suggests that significant improvements in the model accuracy can be achieved by adding expert-guided features to the base features. This improvement in accuracy cannot be reproduced by increasing the model complexity or adding more machine-derived features.

It is also clear that both workflows with machine-derived features and expert-guided machine-derived features are capable of producing robust and reliable SP models with high prediction and cross validation performance. In the case of models using machine-derived features, the monomial-augmented feature set model consistently perform better than the basic set of features. This situation is expected given that simple linear models are insufficient for capturing the complex relationship between low dimensional microstructure features and the short circuit current (J_{sc}) property.

Conclusions

We show using an open-source dataset that feature selection is a valuable approach to constructing SP maps, and that—for a rich-enough feature set—any feature selection approach can be used to train a good SP model. However, machine-only feature engineering and selection methods alone (without human intervention) do not offer the optimal solution to salient feature identification. Our results demonstrate the value of expert knowledge embedded during the featurization step of the structure-property map construction. We have demonstrated that for the application studied here (organic solar cells) SP models employing a combination of machine-learning methods and expert knowledge can achieve similar performance to a time-consuming, entirely hand-crafted SP model used in previous work. This makes the case for development of principled approaches that incorporate expert knowledge into the featurization step during the construction of SP maps.

Acknowledgments

This work was supported by National Science Foundation (1906344 and 1910539). BG acknowledges support from the ONR MURI ONR N00014-19-12453. OW and HL acknowledge the support provided by the Center for Computational Research at the University at Buffalo. BY and SK acknowledge support from NIST 70NANB18H039 (program manager Dr. James Warren).

[6] Here, we use the monomial function of order two. See Supplementary Information for more results.

Data availability

The source code for analysis are available in github: https://github.com/owodolab/FeatureEngineeringOPV.

Declarations

Conflict of interest

The authors declare that they have no conflict of interest.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1557/s43579-021-00147-4.

Nomenclature

$\mathcal{X}$	Raw dataset with microstructures
$\widehat{d}_i$	Salient feature
$\widehat{d}$	Vector of salient features
$\widehat{d}^E$	Vector of salient features derived by the expert
$\widehat{d}'^{\mathrm{mRMR}}$	Vector of salient features derived using the mRMR method from d'
$\widehat{d}^{\mathrm{FS}}$	Vector of salient features derived via forward selection from d.
$\widehat{d}^{\mathrm{mRMR}}$	Vector of salient features derived using the mRMR method from d.
$\widehat{d}^{\mathrm{RF}}$	Vector of salient features derived via random forest method from d.
$\widehat{f}$	Salient features derived via PCA from f
$\widehat{f}'$	Salient features derived via PCA from f'
$\widehat{x}$	Salient features derived via PCA from X
$\widetilde{d}$	Vector of monomial-augmented salient features
$\widetilde{d}'^{\mathrm{mRMR}}$	Vector of monomial-augmented $\widehat{d}'^{\mathrm{mRMR}}$
$\widetilde{d}^{\mathrm{mRMR}}$	Vector of monomial-augmented $\widehat{d}^{\mathrm{mRMR}}$
$\widetilde{f}$	Monomial-augmented $\widehat{f}$
$\widetilde{f}'$	Monomial-augmented $\widehat{f}'$
$\widetilde{M}'_4$	SP model mapping $\widetilde{f}'$ to J_{sc}
$\widetilde{M}_4$	SP model mapping $\widetilde{f}$ to J_{sc}
$\widetilde{M}_1$	SP model mapping $\widetilde{d}'^{\mathrm{mRMR}}$ to J_{sc}
A	Influence coefficients capturing the SP map
d	Vector of descriptors
d'	Expert-enriched vector of descriptors
d_i	Descriptor/feature
F_i	Autocorrelation array of microstructure X_i
F	Function mapping the salient features to property
F'_i	Array of microstructure X_i auto-correlation with state enriched by the expert knowledge
J_{sc}	The short circuit current ($\mathrm{A/m^2}$)
$m(s)$	A state of the microstructure at the location s in X
M_1	SP model mapping $\widehat{d}^{\mathrm{mRMR}}$ to J_{sc}
M'_1	SP model mapping $\widehat{d}'^{\mathrm{mRMR}}$ to J_{sc}
M_2	SP model mapping $\widehat{d}^{\mathrm{FS}}$ to J_{sc}
M'_2	SP model mapping $\widehat{d}'^{\mathrm{FS}}$ to J_{sc}
M_3	SP model mapping $\widehat{d}^{\mathrm{RF}}$ to J_{sc}
M'_3	SP model mapping $\widehat{d}'^{\mathrm{RF}}$ to J_{sc}
M_4	SP model mapping the salient features derived using low dimensional embedding of f.
M'_4	SP model mapping the salient features derived using low dimensional embedding of f'
M_5	SP model mapping $\widehat{x}$ to J_{sc}
M_E	SP model derived by the expert
N	Total number of microstructures in $\mathcal{X}$
P	Material property of interest, here J_{sc}
PC_i	Principal component
q	Order of monomial functions
RL0	Representation layer zero: raw data
RL1	Representation layer one: input features
RL2	Representation layer two: salient features
X	Microstructure data point in $\mathcal{X}$
$\mathcal{X}^f$	Dataset featurized using machine-derived approach
$\mathcal{X}^{f'}$	Dataset featurized using machine-derived approach and enriched with expert knowledge
S	Size of salient features vector
$\widetilde{S}$	Size of salient extended features vector

References

1. B.S.S. Pokuri, S. Ghosal, A. Kokate, S. Sarkar, B. Ganapathysubramanian, npj Comput. Mater. **5**(1), 1 (2019)
2. B.L. DeCost, E.A. Holm, Comput. Mater. Sci. **110**, 126 (2015)
3. S.R. Kalidindi, Int. Mater. Rev. **60**(3), 150 (2015)
4. A. Çeçen, T. Fast, E. Kumbur, S. Kalidindi, J. Power Sources **245**, 144 (2014)
5. R. Bostanabad, Y. Zhang, X. Li, T. Kearney, L.C. Brinson, D.W. Apley, W.K. Liu, W. Chen, Prog. Mater. Sci. **95**, 1 (2018)
6. H. Xu, Y. Li, C. Brinson, W. Chen, J. Mech. Des. **136**(5), 051007 (2014)
7. S. Torquato, Annu. Rev. Mater. Res. **32**(1), 77 (2002)
8. Y. Jiao, F.H. Stillinger, S. Torquato, Proc. Natl. Acad. Sci. USA **106**(42), 17634 (2009)
9. S. Yu, C. Wang, Y. Zhang, B. Dong, Z. Jiang, X. Chen, W. Chen, C. Sun, Sci. Rep. **7**(3752) (2017)
10. M. Teubner, Europhys. Lett. (EPL) **14**(5), 403 (1991)
11. D.M. Dimiduk, E.A. Holm, S.R. Niezgoda, Integr. Mater. Manuf. Innov. **7**(3), 157 (2018)
12. O. Wodo, J. Zola, B.S.S. Pokuri, P. Du, B. Ganapathysubramanian, Process–structure–property map for organic solar cells (2021). https://doi.org/10.5281/zenodo.5061951
13. O. Wodo, B. Ganapathysubramanian, J. Comput. Phys. **230**(15), 6037 (2011)
14. H.K. Kodali, B. Ganapathysubramanian, Model. Simul. Mater. Sci. Eng. **20**(3), 035015 (2012)
15. O. Wodo, J. Zola, B.S.S. Pokuri, P. Du, B. Ganapathysubramanian, Mater. Discov. **1**, 21 (2015)
16. X.Y. Lee, J.R. Waite, C.H. Yang, B.S.S. Pokuri, A. Joshi, A. Balu, C. Hegde, B. Ganapathysubramanian, S. Sarkar, Nat. Comput. Sci. **1**(3), 229 (2021)
17. C.C. Aggarwal, A. Hinneburg, D.A. Keim, *International Conference on Database Theory* (Springer, Berlin, 2001), pp. 420–434
18. O. Wodo, S. Tirthapura, S. Chaudhary, B. Ganapathysubramanian, Org. Electron. **13**(6), 1105 (2012)
19. `GraSPI`: an extensible software for graph-based morphology quantification in organic electronics (2021). https://github.com/owodolab/graspi
20. D. Wheeler, D. Brough, A. Shanker, B. Yucel, S. Voigt, A. Rossi, A. Cecen, F. Hohman, N. Paulson, A. Lohse, A. Medford, aiskakov, S. Kalidindi, A. Castillo, M. Diehl, A. Blekh, M. Whitley, R. Cimrman, E. Popova, S. Mohan, materialsinnovation/pymks: version 0.4.1a1 (2021). https://doi.org/10.5281/zenodo.5043652

21. D.B. Brough, D. Wheeler, S.R. Kalidindi, Integr. Mater. Manuf. Innov. **6**(1), 36 (2017)
22. A. Cecen, T. Fast, S. Kalidindi, Integr. Mater. Manuf. Innov. **5**, 1 (2016)
23. B. Yucel, S. Yucel, A. Ray, L. Duprez, S. Kalidindi, Integr. Mater. Manuf. Innov. **9**, 240 (2020)
24. S. Kalidindi, A. Khosravani, B. Yucel, A. Shanker, A.L. Blekh, Integr. Mater. Manuf. Innov. **8**, 441 (2019)
25. D.T. Fullwood, S.R. Niezgoda, B.L. Adams, S.R. Kalidindi, Prog. Mater. Sci. **55**(6), 477 (2010)
26. A. Gokhale, A. Tewari, H. Garmestani, Scr. Mater. **53**(8), 989 (2005)
27. S.R. Kalidindi, *Hierarchical Materials Informatics: Novel Analytics for Materials Data* (Elsevier, Amsterdam, 2015)
28. B. Ganapathysubramanian, N. Zabaras, Finite Elem. Anal. Des. **44**(5), 298 (2008)
29. R. Olivares-Amaya, C. Amador-Bedolla, J. Hachmann, S. Atahan-Evrenk, R.S. Sanchez-Carrera, L. Vogt, A. Aspuru-Guzik, Energy Environ. Sci. **4**(12), 4849 (2011)
30. G. Chandrashekar, F. Sahin, Comput. Electr. Eng. **40**(1), 16 (2014)