Noname manuscript No.
(will be inserted by the editor)

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

Hao Liu · Berkay Yucel · Daniel Wheeler · Baskar Ganapathysubramanian · Surya Kalidindi · Olga Wodo

Received: date / Accepted: date

Abstract Data-driven approaches now allow for systematic mapping of microstructure with properties. In particular, we now have diverse computational 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. This suggests that feature (down) selection is a good strategy to reduce the complexity of the calibrated model, and any method for feature selection is viable. Moreover, the addition of expert-selected features can significantly improve model performance, emphasizing the need for approaches that assimilate domain knowledge with data-driven approaches.

Keywords Microstructure informatics · machine learning · structure-property map · organic solar cells

H. Liu and O. Wodo
Materials Design and Innovation Department, University at Buffalo, NY, USA +1-716-6451377, E-mail: olgawodo@buffalo.edu

B. Yucel and S. Kalidindi
The School of Materials Science and Engineering, the School of Computational Science and Engineering, Georgia Institute of Technology, GA, USA

D. Wheeler
The National Institute of Standards and Technology, Gaithersburg, MD, USA

B. Ganapathysubramanian
Mechanical Engineering Department, Iowa State University, IA, USA

1 Introduction

The holy grail of materials science is to discover the mechanism governing the properties and describe them in terms of a small set of physically meaningful salient features. It is typically achieved through construction of structure-properties (SP) relationships: $P_i = F(\{d_j\})$, where P_i are the properties of interest, and d_j are the set of salient features or descriptors. The function F is constructed via hypothesis-driven experiments or, more recently, via machine learning approaches [19,8].

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 degree of freedom (or salient features) governing the SP maps. This is because microstructural imaging aims to provide detailed, high-resolution maps. Hence, imaging techniques inevitably produce high-dimensional data sets, while the goal of establishing quantitative SP maps is to derive the smallest set of features that explain the data distribution. Moreover, this set may not be known *a priori* especially for complex multi-physics phenomena governing the properties. Thus, the construction and exploration of F critically depend on the availability of a large set of computable features (we use the words "features" and "descriptors" interchangeably), along with principled approaches of selecting the most informative (or salient) features.

In the case of microstructures, there exist several distinct approaches to "featurize" the microstructure. That is, convert the microstructure (typically an image in 2D or image stack in 3D) into a data representation that one can subsequently perform computations upon. Several such representations have been proposed,

including voxel-based representations, characterization via physical descriptors [5] [25], statistical functions [21, 13], spectral density functions (SDF) [26,20] and machine learning methods [19,8]. For a detailed comparative discussion of various representations, we refer the reader to recent review papers [5,9].

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

We utilize a problem of constructing structure- property maps 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 microstructure- property maps for this application. We utilize an open-source dataset of microstructures and OPV properties – specifically, short circuit current J_{sc} – as our canonical dataset. We utilize a human-derived set of descriptors and machine-derived statistical functions (i.e., 2-point correlation function) and couple these representations with machine learning concepts - feature selection and feature engineering - to construct structure-property (SP) maps. Our reference model for comparison is a SP model carefully derived using expert-derived features. We compare the performance of machine learning approaches with this model and 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 of microstructure featurization, feature selection, and engineering to construct structure-property maps.

2 Materials and methods

This work aims to construct SP map for OPV application. The performance of the OPV device is characterized in terms of its short circuit current (J_{sc}) using a physics-based computational model [16]. The short circuit current is our property of interest and the expected value for our machine learning models to determine their accuracy. The models are calibrated from our dataset of $\sim$ 1700 OPV microstructures generated using Cahn-Hilliard equation [22]. Each microstructure is a two-phase microstructure of size 401×101 pixels and is annotated by one property (J_{sc}). Each microstructure (based on material properties for a well-studied material system, P3HT: PCBM blend) consists of two phases corresponding to electron-donating material and electron-accepting material. The microstructure is sandwiched between two electrodes: an anode and a cathode. See Supplementary Information for more details on data generation and the computational models.

Our reference model, which we denote as M_E, is a surrogate model of a physics-based computational model. The model is based on prior work [24] where experts first defined features and then identified three key features using manual trial and error correlation studies. The model M_E is derived from the same data used in this paper (see details below). We compare the performance of this model against SP maps built with different data-driven featurization schemes. However, the accuracy of models is computed with respect to the results of the physics-based model, as all models are the surrogate models of the physics-based model learned from the data.

Three levels of microstructure representations: Formally, we consider three microstructure data representations levels (RL): the raw data (RL0), the featurized data (RL1), the subset of salient features (RL2) - see Figure 1.

Representation layer zero (RL0): The raw data (i.e., image data) constitutes the representation layer zero. The raw data size depends on the resolution and size of the sample, with even modest samples having sizes in the range of 100^2 (2D) to 100^3 (3D). While it is possible to train SP models that directly map raw data to output [19,17], several challenges exist – including the curse of dimensionality [3] 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). Before moving to the next layer, we formally denote the raw dataset as $\mathcal{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ of N microstructures, where microstructure $\mathbf{X}_i$ is represented by a $(n_x \times n_y)$ bitmap, i.e., $\mathbf{X}_i \in \{0,1\}^{n \times m}$, and $\mathbf{X}_i(x, y)$ is a bitmap pixel at position (x, y).

Representation layer one (RL1): This level corresponds to the feature layer, where transformations are applied to the raw data from RL0. Here, we consider two classes of features: human-derived and machine-derived features (see Figure 1). The human-derived features consist of sets of application-specific descriptors [23]. Such descriptors require (low level) input from experts to formulate and compute. While this featurization approach

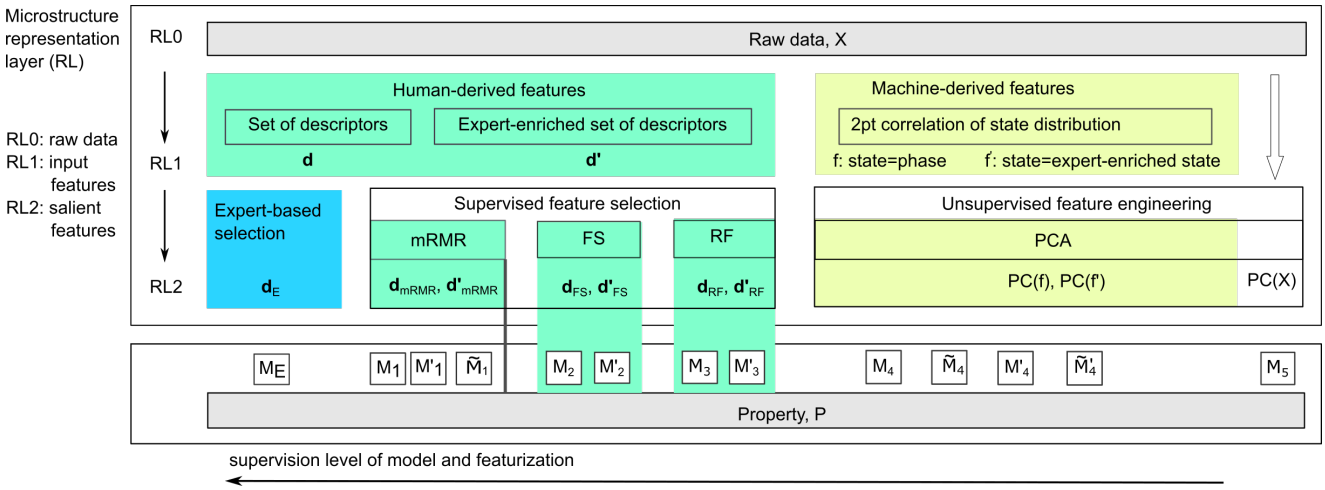

Fig. 1 A taxonomy of microstructure representations with three layers RL0, RL1, RL2 allows principled classification of various approaches to construct microstructure-property maps. RL0 consists of raw data that can be featurized into RL1 using two approaches: human-derived features (here descriptors) and machine-derived features (here two-points correlation function). Both types of features can be enriched with expert knowledge by adding more descriptors or expanding the state for the correlation function. At the third layer - RL3, 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 and are placed below the corresponding salient feature set. Note that all models with a tilde are non-linear, and all models with prime superscripts correspond to the expert enriched feature set.

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, connected components, domain sizes, tortuosity of the paths, and 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 a human-derived set of twenty-one descriptors computed using a graph-based approach [23]. This consists of a basic set of nineteen descriptors, denoted by $\mathbf{d}$, defined based on an understanding of photophysics operations appended with two additional descriptors enriched by the expert. We refer to these two sets as $\mathbf{d} = \{d_1, \ldots, d_l : d_i \in \mathbb{R}\}$ with cardinality of $l = 19$ and the expert-enriched set $\mathbf{d}'$ 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 [1] for a detailed description of these descriptors. In addition to these basic descriptors, two additional descriptors characterizing charge transport are defined. These two additional descriptors were defined 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. The complete set of descriptors is listed in Supplementary Information.

For the machine-derived features, two-point spatial auto-correlations (also known as two-point statistics) are leveraged along with the open source package [2]. There is an extensive literature on using spatial distributions to represent microstructures for structure-property maps [7,27,6,14]. For the two-phase material system under consideration, two auto-correlations and two cross-correlations can be computed. In this work, however, only one the auto-correlation of the electron-accepting phase is used, as the other correlations are redundant, being linearly dependent [10,12]. Let $m(s)$ denote such an array, where s indexes each pixel [1], and m denotes the material/microstructure state of interest at that pixel. In microstructures considered in this work, two phases are considered, and hence m for each pixel can take either zero or one, and auto-correlations are defined as:

$$f(r) = \frac{1}{S_r} \sum_s m(s)m(s+r) \tag{1}$$

where $f(r)$ is the auto-correlation function that reflects the probability of finding the phase of interest at both the tail and the head of a selected vector, r. The factor

[1] a two-dimensional integer index is generally used to represent a two-dimensional image

S_r is the normalizing factor that represents the total number of valid placements of vector r. [2]

Machine-derived features can also be enriched with expert knowledge. The state m initially defined by the phase can be informed by the expert to capture physics-specific characteristics of the microstructure. Here, f' constitutes the expert-enriched feature set following the underlying physics of the application. The property of interest, J_{sc}, is known to depend on the availability of donor-phase adjacent to the anode and acceptor-phase adjacent to the cathode. In light of this, a 1D auto-correlation of material distribution adjacent to the electrode is added to f to generate f'. The expanded state may bias the feature set to include more spatial information in the layer adjacent to the electrodes. We define $f' = \{f, f_{An}, f_{Ca}\}$, where f_{An} and f_{Ca} are 1D auto-correlations on the layers adjacent to the anode and cathode. The remaining steps of the featurization remain identical, with the auto-correlations computed generically.

It is important to note that the 2-point statistics reflect the directional dependencies of the extracted microstructure measures. However, the dimensionality of this representation is fixed and in the same order at the input microstructure. In comparison with descriptors, the statistical functions are generic, and the dimensionality is linked to the size of the input microstructure. However, they lack extensive interpretability. Nevertheless, they can be customized for a particular application by redefining the state (from phase to application relevant quantity) or expanding the state.

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 subsequent SP map. Such concentration of information or salient feature identification is an important step in constructing PS surrogate models. This is because interpolation theory suggests that for a fixed number of samples, more accurate interpolants (from structure features to property) can be constructed when the number of features is smaller [11, 18].

However, identifying a proper set of salient features is challenging. First, there is no uniqueness guarantee for this set of salient features. Different data-driven approaches could result in different sets of features. Second, this set of salient features can be incomplete [3]. Finally, there is no guarantee that this set is 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. Two first two approaches are used on human-derived features, while the last approach is applied on machine-derived features and the raw data.

In the first approach, an expert defines the set of salient features, denoted as $\mathbf{d_E}$. In Figure 1, this approach is marked with the blue box. The set $\mathbf{d_E}$ can be derived using a hypothesis-driven approach or, as in our case, an unsupervised approach relying on the correlation studies. Here, the expert-derived salient features consists of three features $\mathbf{d_E} = \{d_{10}, d_2, min(d_{20}, d_{21})\}$. 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 set of features requires tedious, manual, and time-consuming studies by the domain expert. These three features are used to construct our reference model M_E - in Figure 1 placed below the blue box.

In the second approach, three types of off-the-shelf feature selection techniques are applied: the filter methods, wrapper methods, and embedded methods. For each type of methods, we choose one technique that we briefly describe 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 that enter the SP model. Once the number of features is selected, several models can be constructed as these selection and model construction are independent.

In the wrapper methods, feature selection and machine learning are dependent. Forward selection (FS) is an example method used in this paper, and it involves

[2] In this study, S_r is adjusted to account for the non-periodic nature of the microstructures, see [7,15] for details.

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

the 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 over-fitting. Moreover, for a large pool of descriptors, wrapper methods may be computationally demanding.

Finally, the embedded methods, like the Random Forrest (RF) used in this work, are the most complex feature selection methods. The Random Forest constructs hundreds of decision trees, each of them built the SP map 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 over-fitting 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.

Regardless of the type, feature selection methods belong to 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 Figure 1, we highlight this characteristics by drawing the green background boxes around salient features and the models. In filter methods, the model construction is independent of the feature selection. But the importance score is computed to capture correlation with the property. In Figure 1, we highlight the link with the vertical line. In this paper, mRMR, FS and RF are applied to two sets of descriptors $\mathbf{d}$ and $\mathbf{d}'$ to derive the corresponding features that are denoted with the subscript: $\mathbf{d}_{mRMR}$ and $\mathbf{d}'_{mRMR}$, $\mathbf{d}_{FS}$, $\mathbf{d}'_{FS}$, $\mathbf{d}_{RF}$ and $\mathbf{d}'_{RF}$.

In the third approach - unsupervised feature engineering technique salient features are independent of the properties. Principal Component Analysis (PCA) is an example technique. PCA seeks to reduce the dimensionality of the input data by transforming it into lower dimensional space while capturing high variance in the data and at the same time minimizing the number of dimensions. Since our data $\mathcal{X}_i$ is represented by $(n_x \times n_y)$ bitmap. The total dimensionality is equal to $D = (n_x \times n_y)$. Effectively, PCA finds K largest eigenvectors of the covariance matrix S, where K is the dimensionality presented and $K < D$ [4]. The number of salient features is made based on the residual variance of the eigenvalues of the projected data (or can be selected using the performance of the SP model). Nevertheless, the premise of feature engineering is to identify the small set of features that become the salient features of the data (and not property - hence unsupervised mode). We distinguish between feature selection and engineering, as PCA seeks to transform the input features, effectively redefining (or engineering) the features. In this paper, we apply PCA to the machine-derived features (f and f') from RL1 and the raw data from RL0 ($\mathcal{X}$).

2.1 Microstructure-property maps

Using the salient features from RL2, this study aims to compare and construct the surrogate models of the SP map. The model M_E is considered as the reference model. The form of the model (product of three descriptors) is grounded in the process of the current generation in OPV that involves a sequence of three steps, where the outcome of the subsequent step depends on the previous step. Hence, the property (P or J_{sc}) is modelled as product of three salient descriptors: $P = \prod_i d_E^{A_i}$. We further linearized the model by transforming each descriptor with a logarithmic function. In this way, model $log(P) = \sum_i A_i log(d_i)$ is equivalent to $P = \prod_i d_i^{A_i}$.

This decision guides the form of remaining linear models of this paper: M_1, M'_1, M_2, M'_2, M_3, and M'_3. Note that all models with the prime superscript correspond to the expert-enriched set of descriptors $\mathbf{d}'$, while models without these superscripts correspond to the smaller set of descriptors $\mathbf{d}$. Models with subscripts one, two, and three use the salient features from mRMR, FS, and RF, respectively. In this paper, we also consider the nonlinear model of the polynomial form. All models with tilde sign, for example, $\widetilde{M_1}$ assumes this form. Figure 1 lays out all models below the corresponding set of salient features and methods used to derive them.

We are moving now to unsupervised feature engineering methods. Here, we also construct the linear and nonlinear models from the projected two-point correlation of phase distribution d and expert-enriched state d' into lower-dimensional embedding. The linear model $P = \sum_i^c A_i C_i$, where C_i is the i-th PC score of the low dimensional embedding with c components. The corresponding linear models are M_4, M'_4 and non-linear models $\widetilde{M_4}$, $\widetilde{M'_4}$. Finally, we also construct the model of SP map, M_5, from the raw data again projected to the lower-dimensional embedding.

Table 1 Feature selection and feature engineering applied to two sets of descriptors: d is initial set of descriptors and d' is the expert-enriched set of descriptors with two features defined through time consuming redefinition of descriptors (by the domain expert).

	Expert-derived features and model	Feature selection (mRMR) on human-derived features		Feature engineering (PCA) on machine-derived features (RL1)		PCA on raw data
feature set	d'	d	d'	f	f'	X
salient features	$(d_{10}, d_2, min(d_{20}, d_{21}))$	$(d_3, d_{10}, d_8, d_{15}, d_{19})$	$(d_3, d_{10}, d_{21}, d_8, d_{20})$	$(\mathrm{PC}_1, ..PC_6)$	$(\mathrm{PC}_1, ..PC_6)$	
number	3	5	5	6	6	11 (50)
model	M_0	M_1 ($\widetilde{M}_1$)	$M'_1(\widetilde{M}'_1)$	M_4 ($\widetilde{M}_4$)	M'_4 ($\widetilde{M}'_4$)	M_5
R^2 prediction	**0.97**	0.81 (0.89)	0.94 (0.98)	0.81 (0.94)	0.90 (0.98)	0.60 (0.90)

3 Results and analysis

This study compares the capabilities of various featurization methods (creation, selection, engineering) of the microstructure to enable robust data-driven SP map construction. Altogether, thirteen models of the SP map are construed. Tables 1 contain the summary of selected models, while Figure 2 and 3 depict the results of feature selection and feature engineering methods applied to features from the RL1 to determine the RL2. We report the number of salient features, a list of salient descriptors, the performance measure (R^2) for five-fold validation and prediction and the normalized mean absolute error (NMAE) for prediction. The accuracy is reported with respect to the physics-based computational model.

Our results indicate that model M_E offers the best performance for the linear model. This models is the expert-derived model with only three features manually selected by the expert. The R^2 value for model M_E is 0.97. The superior performance is consistent across training, validation, and prediction. Nevertheless, the model with comparable performance can be constructed for both human-derived and machine-derived features only if features are enriched with expert knowledge and nonlinear model is used ($\widetilde{M}'_1$,$\widetilde{M}'_4$). Below, we compare and contrast various settings of model construction to answer three major questions of this study outlined in the abstract.

We begin by comparing various feature selection methods to construct a SP predictive model. Figure 2 depicts the results for three methods and the associated model M'_1, M'_2, M'_3 for the expert-enriched set of descriptors $\mathbf{d}'$. Three panels of the figure depict the order of descriptors from $\mathbf{d}'$ with the decreasing importance score from top to bottom (except the forward selection where the score is not computed explicitly). Each panel of this figure also contains the insert with reported performance for increasing features with the saturated R^2 guided by the core values. These characteristics are evident from all three panels. For example, the right panel of that figure, where the R^2 increases significantly for the first four features. Moreover, the importance score for these first four features are considerably higher than the following features. The higher values of scores are reflected in the mean $R^2 = 0.94$ of the model for the increasing number of features - as shown in the insert of this Figure. In fact, all three types of feature selection methods (mRMR, FS, RF) offers comparable performance of $R^2 = 0.93$ to 0.95 (see models M'_1, M'_2, M'_3 in Table 2 of the SI). Interestingly, these three models selected a similar number of salient features: four or five. Moreover, the salient feature/descriptors in three sets (see Table 1) are consistent. The interfacial area (d_3), the fraction of donor within 10nm distance to the interface (d_{11}), and the contact area with the interface (d_{20}) 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 Figure 1 in the SI). This demonstrates that 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 involvement from the domain expert. Moreover, it affirmatively answers the question can a small set of features be selected to train a predictive model of SP.

Nevertheless, it should be kept in mind that the features used to construct these models include two expert-crafted features. 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 $\mathbf{d}$. For the same number of features, the performance decreases consistently across three types of methods (mRMR, FS, RF). The R^2 of prediction decreased by 0.1 for forward selection, 0.13 for mRMR, and 0.17 for random forest regression model. Moreover, the performance behavior for an increasing number of features increases only gradually (see Supplementary Information Figure 3, 5 and 6) without a clear saturation behavior. Even with the complete nineteen features, the SP model cannot reach the comparable performance as models with salient features selected from the full set d'. Finally, the lost performance cannot be compen-

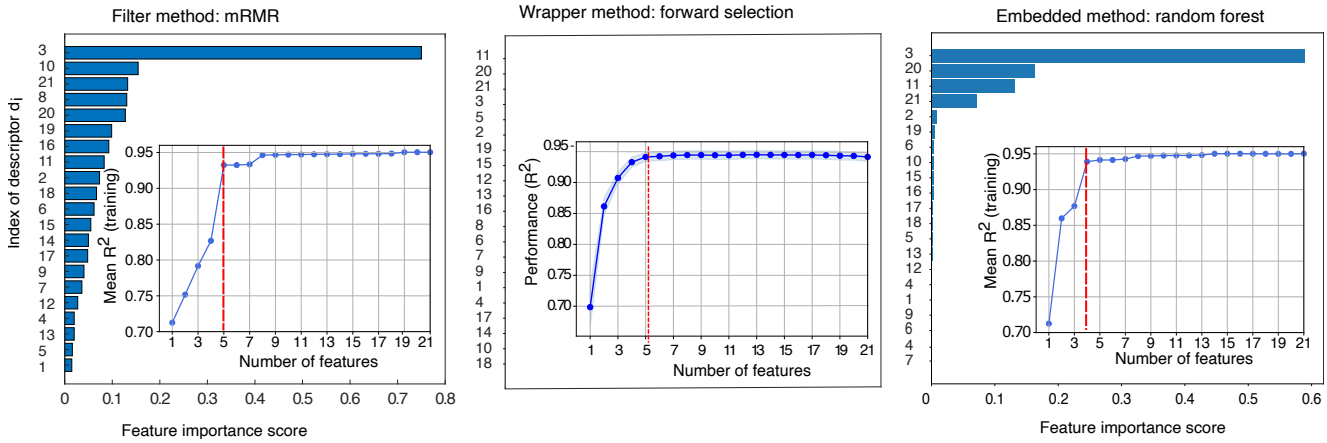

Fig. 2 Performance of feature selection on human-derived features d' and the associated model of SP: (left) filter method (mRMR) and model M_1', (middle) wrapper method (forward selection) and model M_2', and (right) embedded method (random forest) and model M_3'. The red vertical line corresponds to the number of features selected for the model. Note that all methods consistently choose similar salient features (the number and the selected set $d_3, d_{20}, d_{21}, d_{11}$) for which models offer optimal performance.

sated for by increasing the complexity of the model. The R^2 for nonlinear model $\widetilde{M}_1$ the R^2 increases by only 0.10 [4]. These consistent results reiterates the importance of the input features sets, and reinforces 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 that cannot be compensated by the more complex model.

In the second part of this study, four different workflows (M_4,M_4',$\widetilde{M}_4$,$\widetilde{M}_4'$) with machine derived featurization are performed for comparison. Models M_4 and M_4' stand for the linear models with regular machine derived features and expert-enriched machine derived features. $\widetilde{M}_4$,$\widetilde{M}_4'$ stands for non-linear (polynomial) models with regular machine derived features and expert-enriched machine derived features. These structure-property maps are constructed with increasing number of principal component for all workflows described above. Figure 3 depicts the performance of these workflows where y axis represent the prediction (test set) R^2 score and x axis represent the number of PC scores used for the workflow. Our results indicates that all worflows perfom better with increasing number of PC scores (used for model building) up to 10 PC scores. The first 10 PC scores were found to explain >90% of the entire OSC dataset. Therefore, only the first 10 PC scores from each workflow were utilized for establishing structure-property maps with machine-derived features. As seen from the Table 1 and Figure 3, using an expert guidance in machine derived features consistently provide increased prediction performance on both linear and nonlinear models. This observation suggests that additional expert guided features capture valuable information for the correlation to material property, which is not captured by regular machine-derived features (two-point statistics). However, it is also clear that both workflows with machine-derived features and expert-guided machine derived features are capable of producing robust and reliable structure-property maps with high prediction and cross-validation performance.Among the two different modeling strategies for machine-derived features, linear models consistently produced the lowest prediction performance. This situation was expected given the fact that simple linear models are likely to be insufficient for capturing the complex relationship between the low-dimensional microstructure features and the short circuit current (J_{sc}) property.

[4] Here, we use polynomial model of order three, see SI for more results.

4 Conclusion

Our results demonstrate the high value of expert knowledge embedded at any level of the structure-property map construction. Our analysis showcase that it is important to identify the salient features and, more importantly. However, feature engineering and selection methods alone do not offer the optimal solution to salient feature identification. The best performance of the SP maps originates from the expert knowledge infused into the featurization step. In our application - OPV - as supported by the results, when experts knowledge is added to the analysis pipeline, the feature selection and engineering methods offer comparable performance to the manual, time-consuming expert derived SP map in organic solar cells.

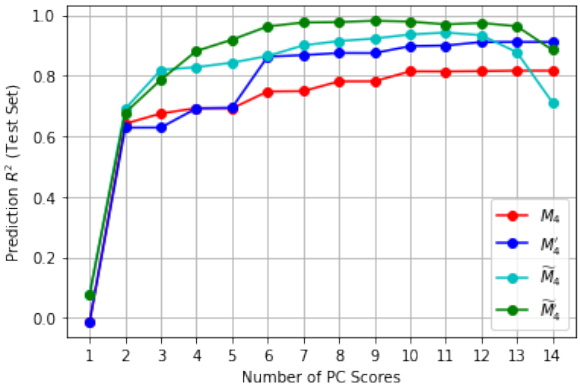

Fig. 3 Feature engineering from machine-derived features used to build model M4: model where microstructure is represented as the two-point correlation of the phase distribution and then transformed to RL2 using PCA with six components being the lower dimensional representation of the structure data.

Acknowledgements This work was supported by National Science Foundation (1906344 and 1910539). BG acknowledges support from the ONR MURI program. OW acknowledges the support provided by the Center for Computational Research at the University at Buffalo.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. `GraSPI`: An extensible software for graph-based morphology quantification in organic electronics. https://github.com/owodolab/graspi (2021)
2. `PyMKS`:the materials knowledge system in python. http://pymks.org/en/stable/ (2021)
3. Aggarwal, C.C., Hinneburg, A., Keim, D.A.: On the surprising behavior of distance metrics in high dimensional space. In: International conference on database theory, pp. 420–434. Springer (2001)
4. Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg (2006)
5. Bostanabad, R., Zhang, Y., Li, X., Kearney, T., Brinson, L.C., Apley, D.W., Liu, W.K., Chen, W.: Computational microstructure characterization and reconstruction: Review of the state-of-the-art techniques. Progress in Materials Science **95**, 1 – 41 (2018)
6. Brough, D.B., Wheeler, D., Kalidindi, S.: Materials knowledge systems in python—a data science framework for accelerated development of hierarchical materials. Integrating Materials and Manufacturing Innovation **6**, 36–53 (2017)
7. Cecen, A., Fast, T., Kalidindi, S.: Versatile algorithms for the computation of 2-point spatial correlations in quantifying material structure. Integrating Materials and Manufacturing Innovation **5**, 1–15 (2016)
8. DeCost, B.L., Holm, E.A.: A computer vision approach for automated analysis and classification of microstructural image data. Computational materials science **110**, 126–133 (2015)
9. Dimiduk, D.M., Holm, E.A., Niezgoda, S.R.: Perspectives on the impact of machine learning, deep learning, and artificial intelligence on materials, processes, and structures engineering. Integrating Materials and Manufacturing Innovation **7**(3), 157–172 (2018)
10. Fullwood, D.T., Niezgoda, S.R., Adams, B.L., Kalidindi, S.R.: Microstructure sensitive design for performance optimization. Progress in Materials Science **55**(6), 477–562 (2010)
11. Ganapathysubramanian, B., Zabaras, N.: A seamless approach towards stochastic modeling: Sparse grid collocation and data driven input models. Finite Elements in Analysis and Design **44**(5), 298–320 (2008)
12. Gokhale, A., Tewari, A., Garmestani, H.: Constraints on microstructural two-point correlation functions. Scripta Materialia **53**(8), 989–993 (2005)
13. Jiao, Y., Stillinger, F.H., Torquato, S.: A superior descriptor of random textures and its predictive capacity. Proceedings of the National Academy of Sciences **106**(42), 17634–17639 (2009)
14. Kalidindi, S., Khosravani, A., Yucel, B., Shanker, A., Blekh, A.L.: Data infrastructure elements in support of accelerated materials innovation: Ela, pymks, and matin. Integrating Materials and Manufacturing Innovation **8**, 441 – 454 (2019)
15. Kalidindi, S.R.: Hierarchical materials informatics: novel analytics for materials data. Elsevier (2015)
16. Kodali, H.K., Ganapathysubramanian, B.: Computer simulation of heterogeneous polymer photovoltaic devices. Modelling and Simulation in Materials Science and Engineering **20**(3), 035015 (2012)
17. Lee, X.Y., Waite, J.R., Yang, C.H., Pokuri, B.S.S., Joshi, A., Balu, A., Hegde, C., Ganapathysubramanian, B., Sarkar, S.: Fast inverse design of microstructures via generative invariance networks. Nature Computational Science **1**(3), 229–238 (2021)
18. Olivares-Amaya, R., Amador-Bedolla, C., Hachmann, J., Atahan-Evrenk, S., Sanchez-Carrera, R.S., Vogt, L., Aspuru-Guzik, A.: Accelerated computational discovery of high-performance materials for organic photovoltaics by means of cheminformatics. Energy & Environmental Science **4**(12), 4849–4861 (2011)
19. Pokuri, B.S.S., Ghosal, S., Kokate, A., Sarkar, S., Ganapathysubramanian, B.: Interpretable deep learning for guided microstructure-property explorations in photovoltaics. npj Computational Materials **5**(1), 1–11 (2019)
20. Teubner, M.: Level surfaces of gaussian random fields and microemulsions. Europhysics Letters (EPL) **14**(5), 403–408 (1991)
21. Torquato, S.: Statistical description of microstructures. Annual Review of Materials Research **32**(1), 77–111 (2002)
22. Wodo, O., Ganapathysubramanian, B.: Computationally efficient solution to the cahn–hilliard equation: Adaptive implicit time schemes, mesh sensitivity analysis and the 3d isoperimetric problem. Journal of Computational Physics **230**(15), 6037–6060 (2011)
23. Wodo, O., Tirthapura, S., Chaudhary, S., Ganapathysubramanian, B.: A graph-based formulation for computational characterization of bulk heterojunction morphology. Organic Electronics **13**(6), 1105–1113 (2012)

24. Wodo, O., Zola, J., Pokuri, B.S.S., Du, P., Ganapathysubramanian, B.: Automated, high throughput exploration of process–structure–property relationships using the mapreduce paradigm. Materials discovery **1**, 21–28 (2015)
25. Xu, H., Li, Y., Brinson, C., Chen, W.: A Descriptor-Based Design Methodology for Developing Heterogeneous Microstructural Materials System. Journal of Mechanical Design **136**(5) (2014). 051007
26. Yu, S., Wang, C., Zhang, Y., Dong, B., Jiang, Z., Chen, X., Chen, W., Sun, C.: Design of non-deterministic quasi-random nanophotonic structures using fourier space representations. Sci Rep **7**(3752) (2017)
27. Yucel, B., Yucel, S., Ray, A., Duprez, L., Kalidindi, S.: Mining the correlations between optical micrographs and mechanical properties of cold-rolled hsla steels using machine learning approaches. Integrating Materials and Manufacturing Innovation **9**, 240 – 256 (2020)

Nomenclature

$\mathbf{d}$	Set of descriptors
$\mathbf{d}'$	Expert-enriched set of descriptors
$\mathbf{X}$	Microstructure data point in $\mathcal{X}$
$\mathcal{X}$	Raw dataset with microstructures
D	Dimensionality of each microstructure
d_j	Descriptor/feature
F	Function to map properties to the salient features
f	Auto-correlation function of microstructure with state being the phase
f'	Auto-correlation function of microstructure with state enriched by the expert knowledge
J_{sc}	The short circuit current (A/m^2)
K	Dimensioality of the low dimensional embedding
$m(s)$	A state of the microstructure at the location s in $\mathbf{X}$
M_E	Expert-derived SP model
N	Total number of microstructures in $\mathcal{X}$
P_i	Materials properties
S	Covariance matrix of $\mathcal{X}$
SP	Structure-property