A Machine Learning-Based Design Representation Method for Designing Heterogeneous Microstructures

Hongyi Xu
Department of Mechanical Engineering, Northwestern University, Evanston, IL 60208
e-mail: hongyixu2014@u.northwestern.edu

Ruqian Liu
Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208
e-mail: rosanne@northwestern.edu

Alok Choudhary
Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208
e-mail: choudhar@eecs.northwestern.edu

Wei Chen
Department of Mechanical Engineering, Northwestern University, Evanston, IL 60208
e-mail: weichen@northwestern.edu

1 Corresponding author.
Contributed by the Design Automation Committee of ASME for publication in the Journal of Mechanical Design. Manuscript received August 21, 2014; final manuscript received February 3, 2015; published online March 5, 2015. Assoc. Editor: Carolyn Seepersad.

1 Introduction

Trial-and-error procedures are the traditional way of material design, which has been mostly guided by experiences and heuristic rules in materials classification, selection, and property predictions. Applying heuristic rules to existing materials databases for searching combinations of processing procedure or material constituents [1,2] is time-consuming and resource intensive; however, microstructure information is often not considered in this process. Constituent-based design approach has relied on heuristic search to choose proper material compositions from materials databases [3,4], but this approach no longer suffices in designing complex microstructural materials systems. To fully explore the potential of computational material design and accelerate the development of advanced materials, “microstructural-mediated design of materials” [5,6] has gained more attention. With this new paradigm, materials are viewed as a complex structural system that has design degrees of freedom in choices of composition, phases, and microstructure morphologies, which can be optimized for achieving superior material properties. In particular, the morphology of microstructure (i.e., the spatial arrangements of local microstructural features) has a strong impact on the overall properties of a materials system. Taking polymer nanocomposites as an example, microstructure percolation determines the electrical conductivity, and the quantity of fillers’ surface area determines the damping properties [7]. Furthermore, heterogeneity in microstructure is the root cause of material randomness at multiple length scales.

There are two major categories of methods: correlation functions and physical descriptors, for quantifying the morphology and heterogeneity of microstructures (also known as “statistical characterization”). The microstructure information of heterogeneous materials can be accurately captured via N-point correlation functions [8–11]. As a balance between computational cost and accuracy, the two-point correlation function (autocorrelation) [12] is widely adopted in practice. However, correlation functions lack clear physical meanings. It is inconvenient to design an optimal correlation functions as they are infinite dimensional [13,14]. Furthermore, correlation function-based microstructure reconstructions are either computationally expensive (when using the moving optimization algorithm [15]), or lacking of stochasticity (when using the phase recovery algorithm [16]). With the physical descriptor-based approach, microstructures are represented by physically meaningful structural parameters (descriptors), such as volume fraction, particle number, and particle size. In our recent research [13], we classified microstructure descriptors into three categories: composition, dispersion, and particle size. In our recent research [13], we classified microstructure descriptors into three categories: composition, dispersion, and particle size. The major strengths of physical descriptors are the clear physical meanings they offer and meaningful mappings to processing parameters [17]. We have developed a descriptor-based methodology for characterization and reconstruction of polymer nanocomposites [14,18]. However, the descriptors were chosen based on experiences. A systematic approach of identifying key microstructure descriptors as material design variables is needed.

Material informatics [19,20] is a growing area that leverages information technology and data science to represent, parse, store, manage, and analyze the material data. The goal is to share and mine the data for uncovering the essence of materials, and
accelerate the new material discovery and design [21]. Data mining and machine learning techniques have been applied to exploit material databases and discover trends and mathematical relations for material design. To manage the information complexity of using large-dimensional representations of microstructures, recent work has attempted unsupervised microstructure dimensionality reduction via manifold learning [9] and kernel principal components [22]. However, dimension reduction of microstructure parameters considering the microstructure only does not reflect its impact on material properties of interest so that the reduced parameter set does not address the direct need of material design. Supervised learning [23], a concept in machine learning where labeled training data are used to infer a relationship, has been employed in establishing the process–composition–property relation for metals [1,2] and predicting polymer composites’ properties based on the composition–property database [24–26]. However, limited efforts have been made on modeling the microstructure–property relation using statistical learning and further reducing the high dimensionality of microstructure representations obtained from analyzing microscopic images.

High dimensionality is handled in machine learning by feature selection and extraction, to reduce the number of variables in a system by either selecting a subset of relevant features, or transforming the original high-dimensional feature space into a space of fewer dimensions. Both selection and extraction can be either supervised or unsupervised. The transformation incurred by extraction methods usually refers to a linear or nonlinear combination of the original variables, in order to construct new features for improved description of data. In this regard, extraction methods are not suitable for our needs. We rather want to retain the clear physical meanings of features (descriptors) so as to use them as design variables. Feature selection, on the other hand, chooses a subset of more informative features from the original set and well fits our scenario. Only looking at the microstructure descriptors forms an unsupervised learning process. If the corresponding responses (behavior) of microstructures, in our case, the morphology and properties, are also available, supervised learning provides more insights in the selection process.

Existing supervised feature selection methods typically involve developing heuristics or measures to evaluate the worth of features. Examples of heuristics developed in literature include information gain [27], Gini index [28], Chi-square, and other distance measures. The limitation is that they can only handle discrete variables as the supervisory signal, as when the desired output is within a set of a small number of known labels. However, both microstructure correlation functions and properties in our case are provided as continuous values, and therefore pose challenges for the feature selection procedure. What’s more, distance measure based heuristics do not take into account the feature interactions and dependencies, for example, the surface area of filler phase and that of matrix phase in our descriptor group have a high dependency, which cannot be appropriately addressed by common distance measures. The family of Relief algorithms, beginning with the basic form of Relief [29] and being later adapted into ReliefF [30] and RReliefF, are efficient and effective heuristic measures that correctly estimate the quality of features considering their capability of differentiating opposite-class training examples. The first two in the family are developed for discrete problems. RReliefF [31], the algorithm employed in this research, accounts particularly for continuous problems. For simplicity, we refer it as Relief.

In this paper, we propose a four-step machine learning methodology for identifying the key microstructure descriptors as potential material design variables. In step 1, image analysis is applied to gather an initial set of potential microstructure descriptors (Sec. 2), to understand the dependencies among the descriptors and the topological constraints of the microstructure morphology (Sec. 3.1). In step 2, an image analysis-based supervised learning further reduces the descriptor set by analyzing each descriptor’s influence on the microstructure morphologies represented by the correlation functions (Sec. 3.2). In step 3, material property-based supervised learning is employed using data obtained from physics-based simulations or from literature for further dimension reduction to identify the key set of descriptors (design variables) that have the largest impact on properties of interest (Sec. 3.3). In step 4, microstructure design variables are selected from key descriptors (Sec. 3.4) by maximizing the impact score and minimizing the dependency. We demonstrate the strength of the proposed method with the design of polymer nanocomposites (Sec. 4).

2 Technical Background of Statistical Microstructure Representations

In this section, we provide the technical background of microstructure characterization with the example of biphase nanoparticle-reinforced polymer composite. Statistical microstructure characterization enables a quantitative understanding of the microstructure–property relationship. Two types of microstructure characterization techniques are introduced: correlation function-based method (Sec. 2.1) and descriptor-based method (Sec. 2.2). We also summarize a list of commonly used descriptors covering composition, dispersion status, and geometry information of the inclusions.

2.1 Correlation Function-Based Microstructure Characterization. A wide range of microscopic imaging techniques such as scanning electron microscopy (SEM) [7,32] and transmission electron microscopy [33] are applicable to obtain the digital microstructure images for statistical characterization. In the step of image preprocessing, the biphase microstructure images are denoised and binarized with the volume fraction of each phase maintained. In the binary image, pixels in the matrix phase are marked by “0” and pixels in the filler phase are marked by “1.” Figure 1 illustrates the transformation of the gray scale SEM

---

**Fig. 1** SEM images of polymer nanocomposites, binary image, and correlation function-based characterization.
image (left) to a binary image (middle), where black pixels represent nanoparticle filler and white pixels represent polymer matrix. The binary pixelated images are used in both correlation function- and descriptor-based characterization.

In this work, we collect four types of correlation functions [34,35]: two-point correlation function (nondirectional two-point autocorrelation functions), (two-point) surface correlation function, lineal path function, and radial distribution function (Fig. 1), which are widely used for an accurate representation of microstructure with affordable computational costs. Correlation functions are functions of distance \( r \). These four correlation functions are complementary to each other, emphasizing on different aspects of microstructure features. Characterization using multiple correlation functions together has been reported in Refs. [15,34,36]. For biphasic heterogeneous polymer nanocomposites' microstructure studied in this work, it is assumed that the aforementioned four correlation functions can adequately describe the filler morphology.

2.2 Descriptor-Based Microstructure Characterization. A descriptor-based approach is proposed in our prior work to represent microstructure morphologies using three levels of microstructure features [14]: composition, dispersion, and geometry (Fig. 2). Composition descriptors distinguish different phases and describe their volume/weight percentage in the material, such as volume fraction of filler in polymer composites. Dispersion status descriptors depict the inclusions' spatial relation and their neighbor status, such as the nearest neighbor distance, number of filler clusters [11,32,37,38], etc. Geometry descriptors are on the lowest length scale, which describe the inclusions’ shapes. Geometry descriptors include the inclusions’ size distribution, surface area, surface-to-volume fraction, roundness, eccentricity, elongation, rectangularity, tortuosity, aspect ratio, etc. [8–10,32,38–42]. The descriptor-based methodology is featured by four strengths: the well-defined physical meaning of microstructure characteristics, the high correlation with material properties, the low computational cost in characterization/reconstruction, and the low dimensionality of parameterized microstructure characteristics that enables parameter-based optimal microstructure design. With a sufficient descriptor set, high orders of microstructure information can be captured [14].

In this paper, we collect a large set of descriptors from literature as candidates of microstructure design variables. This section covers descriptors used in polymer nanocomposites, alloy, fiber composites, ceramic composites, etc. In previous works, different descriptors are chosen for different materials based on expertise. Often times, the descriptors used in a single work only capture the microstructure features that are highly related to the interested properties, while all the other microstructure features are neglected. Therefore, to avoid bias in the key descriptor learning, it is necessary to include a wide range of descriptors from different types of materials. The full candidate descriptor set is referred to as the “full descriptor set” in this paper. The collection of descriptor titles and their definitions are provided in Table 1. There are 17 descriptors in the list, in which each statistical descriptor is represented by four parameters (first to fourth order moments). In total, the 17 microstructure descriptors are represented using 56 descriptor parameters.

### Table 1 Collected microstructure descriptors. Statistical information includes first to fourth orders of moments: mean, variance, skewness, and kurtosis.

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Definition</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Composition</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VF</td>
<td>Volume fraction</td>
<td>Deterministic</td>
</tr>
<tr>
<td>Dispersion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( r_{\text{rad}} )</td>
<td>Cluster’s nearest surface distance</td>
<td>Statistical</td>
</tr>
<tr>
<td>( \theta )</td>
<td>Principle axis orientation angle [43]</td>
<td>Statistical</td>
</tr>
<tr>
<td>( I_{\text{filler}} )</td>
<td>Surface area of filler phase</td>
<td>Deterministic</td>
</tr>
<tr>
<td>( I_{\text{matrix}} )</td>
<td>Surface area of matrix phase</td>
<td>Deterministic</td>
</tr>
<tr>
<td>( N )</td>
<td>Cluster number</td>
<td>Deterministic</td>
</tr>
<tr>
<td>( V_{\text{VF}} )</td>
<td>Local VF of Voronoi cells [39]</td>
<td>Statistical</td>
</tr>
<tr>
<td>Geometry</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( r_p )</td>
<td>Pore sizes (inscribed circle’s radius) [44]</td>
<td>Statistical</td>
</tr>
<tr>
<td>( A )</td>
<td>Area</td>
<td>Statistical</td>
</tr>
<tr>
<td>( r_c )</td>
<td>Equivalent radius, ( r_c = \sqrt{A/\pi} )</td>
<td>Statistical</td>
</tr>
<tr>
<td>( \delta_{\text{cmp}} )</td>
<td>Compactness [45]</td>
<td>Statistical</td>
</tr>
<tr>
<td>( \delta_{\text{md}} )</td>
<td>Roundness [46]</td>
<td>Statistical</td>
</tr>
<tr>
<td>( \delta_{\text{ec}} )</td>
<td>Eccentricity [46]</td>
<td>Statistical</td>
</tr>
<tr>
<td>( \delta_{\text{ap}} )</td>
<td>Aspect ratio [33,41]</td>
<td>Statistical</td>
</tr>
<tr>
<td>( \delta_{\text{re}} )</td>
<td>Rectangularity [46]</td>
<td>Statistical</td>
</tr>
<tr>
<td>( \delta_{\text{tor}} )</td>
<td>Tortuosity [46]</td>
<td>Statistical</td>
</tr>
</tbody>
</table>

3 Machine Learning-Based Identification of Key Descriptors

In the presence of a large number of microstructure descriptors, the key research questions is how to represent the microstructural design space quantitatively using a descriptor set that is sufficient yet small enough to be tractable. A four-step machine learning-based method is proposed to exploit the microstructure–property database (Fig. 3). The four steps include: (1) elimination of redundant descriptors using descriptor–descriptor correlation analysis; (2) microstructure correlation function-based supervised learning for further dimension reduction; (3) property-based supervised learning to identify key descriptors; (4) determination of microstructure design variables based on the optimization criteria of maximizing the impact score and minimizing the within-group correlations of the selected descriptor set. Steps 1 and 2 are image analysis-based procedures, which do not require expensive finite element analysis (FEA) simulations. These two steps will provide a fast reduction of the size of a candidate descriptor set. Both steps 2 and 3 involve supervised learning. Step 3 needs structure–property data from either high-fidelity simulations or from literature. Step 4 is an optimization-based descriptor subset selection process.

To build a rich set of data, multiple microstructure images are collected for the type of materials of interest. For each material sample, one representative volume element (RVE) size image or multiple statistical volume element size images should be collected [47]. RVE has spatially invariant properties and microstructural statistics. For each image, a full set of microstructure representations (correlation functions and descriptors) are evaluated using the characterization techniques introduced in Sec. 2.

3.1 Descriptor–Descriptor Correlation Analysis for Identifying Redundant Descriptors. In step 1 of the proposed framework, redundant descriptors are identified by the pair-wise descriptor–descriptor correlation analysis. Some descriptors may be strongly correlated due to the pre-existing relations. For example, geometry descriptor cluster area \( A \) and major radius \( r \) of the fillers in microstructure 1 (Fig. 4) follow a strict mathematical relation of \( A = \pi r^2 \), so these two descriptors \( A \) and \( r \) are
interchangeable to each other; therefore, one of them becomes redundant. However, in microstructure II (Fig. 4), there is no clear mathematical relation between $A$ and $r$, so both descriptors are to be kept for designing microstructure II.

Since the mathematical relation may not necessarily be linear, rank correlation is preferred as the measure of descriptor–descriptor correlation to the widely used correlation coefficient, which only measures the linear dependence between variables. In statistics, rank correlation (Kendall’s $\tau$) measures the degree of similarity between two rankings, and is used to assess the significance of the two variables’ relation. The formula computing Kendall’s $\tau$ is

$$\tau = \frac{a - b}{\frac{1}{2} n(n-1)}$$

where $a$ is the number of concordant pairs, and $b$ is the number of discordant pairs. Rank correlation detects the nonlinear relation between variables. For example, a data set \{x, y\} following $y_i = x_i^2, x_i \in [0, 1]$ has a rank correlation of 1 (which indicate a perfect relation), but has a correlation coefficient smaller than 1.

3.2 Correlation Function-Based Supervised Learning.

Step 1 of the proposed framework eliminates a few redundant microstructure descriptors based on the descriptor–descriptor correlations, but it does not provide any information on the significance of each descriptor to the properties of interest. Supervised learning is needed to search the key descriptors. It is not realistic to directly conduct property-based supervised learning on the large set of microstructure descriptors. The high dimensionality of descriptor set requires a great amount of microstructure samples (e.g., 10 times of the dimensionality) in structure–property simulations for supervised learning. This process may not be affordable due to the high computational costs of simulations. For example, a high fidelity damping property simulation that explicitly models the microstructures of an 80 voxel size 3D microstructure takes over 80 h [48,49]. Therefore, a simulation-free, image analysis-based supervised learning step (step 2) is proposed to further reduce the number of candidate descriptors before property-based supervised learning in step 3.

In step 2, each descriptor’s influence on microstructure morphology is evaluated based on their influences on the four correlation functions introduced in Sec. 2.1. For each descriptor, four impact scores (on four correlation functions) are evaluated using supervised learning algorithm. Their average is taken as the descriptor’s final score. Relief [31] is employed as the supervised learning algorithm, which takes descriptors as input features and the sum of correlation function values as the supervisory signal. We take the sum of first 50 points of correlation functions, which represents the homogenized high-strength correlation within a distance of 50 pixels, 295 nm.

Relief uses a statistical method and avoids heuristic search. Only statistically relevant features are selected. The key idea of the Relief algorithm is to estimate the quality of attributes according to how well their values distinguish between instances that are near to each other within a local context. The pseudo code of the

![Diagram of machine learning-based microstructure descriptor identification](image)

Fig. 3 Framework of machine learning-based microstructure descriptor identification

![Illustration of redundant microstructure descriptors](image)

Fig. 4 Illustration of redundant microstructure descriptors. In microstructure I, area $A$, and major radius $r$ can replace each other; in microstructure II, both are needed for a full microstructure representation.
The process is repeated for \( m \) times, where \( m \) is a user-defined parameter.

For categorical attributes, the outcome of the function diff (\( \text{Attribute} \), Instance1, Instance2) is a binary value, 0 being the values of \( \text{Attribute} \) agree between Instance1 and Instance2 and 1 otherwise. For continuous attributes, the function diff (\( \text{Attribute} \), Instance1, Instance2) is defined as

\[
\text{diff}(A, I_1, I_2) = \frac{|\text{value}(A, I_1) - \text{value}(A, I_2)|}{\max(A) - \min(A)}
\]  

The above function calculates the difference between the values of \( \text{Attribute} \) for two instances, where Instance1 is a random instance, and Instance2 can be either hit H or miss M.

To handle regressional cases, instead of the above difference functions, a kind of probability is introduced to address how much the predicted values of two instances are different. This probability can be modeled with the relative distance between the predicted (class) values of two instances. The output of this algorithm, after going through all instances, is the quality estimation vector (impact factors) \( W \) that represents the estimations of the qualities of each feature.

Finally, according to the obtained quality estimation vector \( W \), features are ranked, and how many are to be selected from the ranked list is a decision subject to the user. Relief requires linear time in the number of given features and the number of instances regardless of the target concept to be learned.

3.3 Property-Based Supervised Learning. The end goal of the machine learning framework is to identify key microstructure descriptors as design variables to optimize for achieving target material properties. In the third step of the framework, supervised learning is employed to study descriptors’ influences on properties of interest. One microstructure image is one “instance” in the learning algorithm. The reduced descriptor set obtained from the first two steps is used as inputs (attributes), and material properties are taken as the supervisory signal. The properties of microstructure samples are either obtained from advanced FEA or collected from literature. The Relief algorithm is employed again to calculate the score of each descriptor on each property of interest. The learning result is normalized such that the scores of all microstructure descriptors are in the range of [0, 1] and add up to 1. If multiple properties are considered in material design, the supervised learning is applied on each property for all descriptors, and then the scores are added together to determine the final ranking of the microstructure descriptors.

3.4 Determination of Microstructure Design Variables. A small set of microstructure descriptors are chosen from the key descriptors as microstructure design variables. It is not realistic to include all key descriptors as design variables because the strong descriptor–descriptor correlations may lead to unrealistic (infeasible) designs. Step 1 of the learning process only eliminates “repetitive” descriptors (descriptors of very strong correlations), so it does not necessarily mean that the descriptors kept after step 1 are independent or weakly correlated. The microstructure design variables should have high contribution to material properties (high ranking from machine learning) and high independency (low descriptor–descriptor correlation). A combinatorial search is conducted to determine the most proper subset of descriptors by formulating the problem as a two-objective heuristic search.

Given the number of design variables \( n \), find descriptors \( d_1, d_2, \ldots, d_n \) such that:

\[
\begin{align*}
\text{Min:} & \sum_{i=1}^{n} c_{ij}, & \text{where } i = 1, 2, \ldots, n, & j = 1, 2, \ldots, n, i \neq j; \\
\text{Max:} & \sum_{i=1}^{n} S_i \\
C_{ij} & \text{is the correlation between any two descriptors, } S_i \text{ is the } i^{th} \text{ descriptor’s contribution to the properties (impact score).}
\end{align*}
\]

4 Design of Polymer Composites Using Reduced Descriptor Set

The addition of reinforcing particles to polymer nanocomposites’ matrix can lead to significant improvements in homogenized mechanical properties even at a very low filler concentration [50]. High impact of the quantity and morphologies of nanoparticle fillers on damping property makes it an interesting design problem for microstructure optimization. This section demonstrates how to use the proposed method to determine the key microstructure descriptors as design variables for carbon black nanoparticle filled polymer elastomers. All material samples collected have the same type of fillers, but are produced under different processing conditions, which directly impact the morphology of nanoparticle clusters in the polymer matrix. Fifty-six microstructure images are collected on materials produced under 11 different processing conditions. The pixel size of the SEM images is 1000 × 1000. The physical size is 5.9 × 5.9 \( \mu m \), which can be considered as RVE. Sample images are shown in Figs. 1 and 3, respectively. The microstructure information includes four types of correlation functions and 17 types of microstructure descriptors discussed in Sec. 2.2 (also see the Appendix). In this paper, we analyze the damping property, which is defined as

\[
\tan\delta(\omega) = \frac{G''(\omega)}{G'(\omega)}
\]

where \( G'' \) is the shear loss modulus (GPa), \( G' \) is the shear storage modulus (GPa). \( \tan\delta, G', G'' \) are all functions of \( \omega \), the frequency of excitation (Hz).

To achieve a long wear life, low rolling resistance, and high wet traction of tire materials, it is desired that the nanocomposites have a \( \tan\delta \) curve with low value in the low frequency domain (smaller than 1 × 10^{-1} Hz), high value in the normal (from 1 × 10^{-1} to 1 × 10^{3} Hz), and high frequency domains (larger than 1 × 10^{3} Hz) [14]. Typically, \( \tan\delta \) is a smooth bell-shaped curve in the frequency domain. It has low values at the two ends and a peak in the middle (Fig. 3). Therefore, we choose three property characteristics as the design criteria: value of the first point on
tan δ (L, representing damping property in low frequency domain), value of the peak on tan δ (P, representing normal frequency domain), and value of the last point on tan δ (H, representing high frequency domain). L, P, and H represent three different desired properties of tire material, respectively: low L leads to low rolling resistance; high P leads to high wet traction; high H leads to low wear. The goal is to achieve good performances on all three properties simultaneously. Thus, the design problem is formulated as a multi-objective optimization problem.

Find a set of microstructure descriptors, s.t.: Min L; Max P (Min −P); Max H (Min −H);

4.1 Results of Correlation-Based Feature Selection. In step 1 of the proposed framework (descriptor–descriptor correlation analysis), the rank correlation (Kendall’s tau) is calculated for each pair of descriptor parameters and written into a 56 × 56 symmetric matrix. The correlation matrix is reordered using Cuthill–Mckee algorithm [51], which permutes the symmetric correlation matrix to obtain a band matrix form with a small bandwidth. The permuted correlation matrix is shown in Fig. 5. The correlation values are represented by colors. Darker color means a higher correlation, and lighter color means a lower correlation. Numbers on the X-matrix is shown in Fig. 5. The correlation values to obtain a band matrix form with a small bandwidth. The per-

algorithm [51], which permutes the symmetric correlation matrix (correlation matrix as shown in Figs. 5(a) and 5(b)), the binarized correlation matrix. The bright pixel means a correlation larger than or equal to the threshold; the dark pixel means a correlation smaller than the threshold. Two thresholds are tested: 0.9 and 0.8.

Fig. 5 (a) The permuted rank correlation matrix shows intracorrelated descriptor groups. Larger correlations are marked by darker colors. White color means the correlation is 0 (no correlation). The sequential numbers on X, Y axis represent different descriptors. Refer to the Appendix for the meanings of sequential numbers. (b) and (c): the binarized correlation matrix. The bright pixel represents correlation values passing the threshold, and the dark pixel represents those failing the threshold. The highly correlated descriptors (Table 2) are considered as interchangeable, as the result, one of each pair can be eliminated to reduce the dimension. With a higher threshold (larger than 0.9 or smaller than −0.9), the dimension is reduced from 56 to 54. Redundancy exists in one pair of dispersion descriptors (I_matrix and I_filler are the measurements of surface area) and one pair of geometry descriptors (A_1 and r_{c_1} describe the size of each filler cluster). When using a lower threshold (larger than 0.8 or smaller than −0.8), the size of the candidate descriptor set can be reduced by 15 (redundancy exists in six pairs of dispersion descriptors and nine pairs of geometry descriptors).

Another important conclusion can be made from correlation analysis. Geometry descriptors tend to be independent with composition and dispersion descriptors. High correlations exist between descriptors of the same category. This observation validates our method of classifying all microstructure descriptors into three categories [14] according to the different levels of details (global versus local).

4.2 Correlation Function-Based Supervised Learning. In step 2, correlation function-based supervised learning identifies the microstructure descriptors that have a high impact on the microstructure morphology. For materials systems of different microstructure features, correlation function-based learning is

| Table 2 The highly correlated descriptors identified with different thresholds |
|-----------------------------|-----------------------------|
| Threshold = 0.9             |                            |
| I_{matrix} ↔ I_{filler}     | Dispersion                 |
| A_1 ↔ r_{c_1}               | Geometry                   |
| Threshold = 0.8             |                            |
| VF ↔ V_{VF_1}               | Composition, dispersion    |
| I_{matrix} ↔ I_{filler}     | Dispersion                 |
| Ø_2 ↔ Ø_4                   | Dispersion                 |
| r_{rad_3} ↔ r_{rad_4}       | Dispersion                 |
| V_{VF_3} ↔ V_{VF_4}         | Dispersion                 |
| A_4 ↔ r_{c_4}               | Geometry                   |
| δ_{rec_3} ↔ δ_{rec_4}       | Geometry                   |
| δ_{rad_1} ↔ r_{c_1}         | Geometry                   |
| A_2 ↔ r_{c_2}               | Geometry                   |
| δ_{as_3} ↔ δ_{as_4}         | Geometry                   |
| δ_{rec_3} ↔ δ_{rec_4}       | Geometry                   |
From the learning results listed in Table 3, we can conclude that dispersed fillers; the dielectric system are featured by low volume descriptors that are identified from two different microstructural I1 dielectric system, the bright part represents fillers; in dielectric system, the dark spots represent fillers.

**Fig. 6 Sample images of damping system (evenly dispersed fillers) and dielectric system (clustering of fillers). In damping system, the bright part represents fillers; in dielectric system, the dark spots represent fillers.**

able to find different sets of significant descriptors. As verification, we compare the ranking and impact scores of significant descriptors that are identified from two different microstructural materials systems: a damping system and a dielectric system. The damping system has high volume fraction (9–30%) and evenly dispersed fillers; the dielectric system are featured by low volume fraction (<4%) and sparsely distributed fillers clusters (Fig. 6). From the learning results listed in Table 3, we can conclude that (1) the geometry descriptors have higher significance in the dielectric system that has fewer fillers; (2) the descriptors have closer impact scores in the evenly dispersed damping system. Meanwhile, volume fraction and surface areas are identified as important descriptors in both systems.

### 4.3 Property-Based Supervised Learning of Key Descriptors.

For the damping system, the results of step 2 correlation function-based learning and step 3 property-based learning are listed in Table 4. In step 2, we calculate the scores of all 56 microstructure descriptors based on their influences on correlation functions. The scores add up to 1. The size of candidate descriptor set is reduced to 20, when a threshold of 0.5 is set on the sum of scores of a reduced descriptor set. The top 20 descriptors are identified as significant descriptors. Next in step 3, property-based learning, is conducted to further evaluate the significant descriptors’ impacts on material properties. The final ranking and scores of the 20 significant descriptors are listed in the right half of Table 4. In addition, this study leads to another two important observations:

1. For the type of materials studied in this paper, the composition and dispersion descriptors have strong influences on both correlation function and properties. Composition and dispersion descriptors are on higher (global) levels than the geometry (local) descriptors. The material property of interest (damping property) is the averaged response of the bulk material, so it is expected to be highly correlated with the high level descriptors.

2. For this particular type of materials, we observed a high similarity between the correlation function-based descriptor ranking and the property-based descriptor ranking. The two rankings have the same top 5 descriptors, and share 9 descriptors out of top 10.

The benefits in computational efficiency are concluded from the perspective of problem dimensionality (number of design variables). The size of candidate descriptor set is reduced from 56 to 41 after step 1 (when threshold is 0.8). According to our recommendation that the sample number should be at least 10 times of the number of design variables, it requires 410 microstructure samples (410 high fidelity simulations) when property-based machine learning is directly applied on the 41 descriptors. On contrary, the correlation function-based learning (step 2) eliminate low-impact descriptors. The dimension is further reduced to 20. It means that 210 less samples (210 high fidelity simulations) are required for step 3 property-based supervised learning.

### 4.4 Design Validation: Optimization of Polymer Nano-composites’ Microstructure.

A comparative study of microstruc-

<table>
<thead>
<tr>
<th>Step 2: correlation function-based</th>
<th>Step 3: property-based</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Rank</strong></td>
<td><strong>Descriptor</strong></td>
</tr>
<tr>
<td>1</td>
<td>( I_{\text{filler}} )</td>
</tr>
<tr>
<td>2</td>
<td>( \text{VF} )</td>
</tr>
<tr>
<td>3</td>
<td>( \text{VF} )</td>
</tr>
<tr>
<td>4</td>
<td>( \text{VF}_1 )</td>
</tr>
<tr>
<td>5</td>
<td>( \text{VF}_2 )</td>
</tr>
<tr>
<td>6</td>
<td>( \text{VF}_3 )</td>
</tr>
<tr>
<td>7</td>
<td>( \text{VF}_4 )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Rank</strong></th>
<th><strong>Descriptor</strong></th>
<th><strong>Category</strong></th>
<th><strong>Score</strong></th>
<th><strong>Descriptor</strong></th>
<th><strong>Category</strong></th>
<th><strong>Score</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( I_{\text{filler}} )</td>
<td>Dispersion</td>
<td>0.0362</td>
<td>( I_{\text{matrix}} )</td>
<td>Dispersion</td>
<td>0.0623</td>
</tr>
<tr>
<td>2</td>
<td>( \text{VF} )</td>
<td>Composition</td>
<td>0.0348</td>
<td>( r_{\text{disp}} )</td>
<td>Dispersion</td>
<td>0.0587</td>
</tr>
<tr>
<td>3</td>
<td>( \text{VF} )</td>
<td>Dispersion</td>
<td>0.0335</td>
<td>( \delta_{\text{disp}} )</td>
<td>Geometry</td>
<td>0.0491</td>
</tr>
<tr>
<td>4</td>
<td>( \text{VF}_1 )</td>
<td>Dispersion</td>
<td>0.0286</td>
<td>( \delta_{\text{disp}} )</td>
<td>Geometry</td>
<td>0.0294</td>
</tr>
<tr>
<td>5</td>
<td>( \text{VF}_2 )</td>
<td>Dispersion</td>
<td>0.0263</td>
<td>( \delta_{\text{disp}} )</td>
<td>Geometry</td>
<td>0.0211</td>
</tr>
<tr>
<td>6</td>
<td>( \text{VF}_3 )</td>
<td>Dispersion</td>
<td>0.0246</td>
<td>( \delta_{\text{disp}} )</td>
<td>Geometry</td>
<td>0.0189</td>
</tr>
<tr>
<td>7</td>
<td>( \text{VF}_4 )</td>
<td>Dispersion</td>
<td>0.0243</td>
<td>( \delta_{\text{disp}} )</td>
<td>Geometry</td>
<td>0.0082</td>
</tr>
</tbody>
</table>

Journal of Mechanical Design
MAY 2015, Vol. 137 / 051403-7
feasibility). However, they are not necessarily the best choice for controlling the properties of interest. One similar example of empirical choice of three descriptors is reported in Ref. [52], where the assumption of spherical fillers is made. It is a challenging task to check all possible descriptor combinations in microstructure optimization and illustrate that our proposed descriptor set is the best choice. Our objective is to demonstrate that the key descriptors identified by the machine learning techniques can achieve better material properties than the designs obtained from the descriptors identified using experts’ knowledge. The value ranges of the descriptors are listed as follows. VF and N are deterministic; other descriptors’ means are considered.

Design variable set 1 (statistical learning):

\[ VF \in [0.1, 0.3]; \quad N \in [100, 300]; \quad \delta_{\text{rd}} \in [1, 4] \]

Design variable set 2 (empirical):

\[ VF \in [0.1, 0.3]; \quad r_{\text{rd}} \in [10, 40]; \quad \delta_{\text{ec}} \in [\pi/4, 1] \]

We show that the statistically learned microstructure descriptors can capture microstructure information of the studied materials accurately. The descriptor-based microstructure reconstructions are compared with the original image (Fig. 7). Both reconstructions using statistically learned descriptor set and empirical descriptor sets, respectively, match well with the original image in visual comparison and correlation functions. The statistically learned descriptor set is more accurate, as it has a smaller sum of squared error in two-point correlation function compared with the reconstruction from empirical descriptor set (1.83 \times 10^4 versus 6.26 \times 10^4).

Microstructure optimization is conducted using a DOE/metamodeling-based optimization strategy. For both descriptor sets 1 and 2, design of experiment (DOE) is applied to explore the design space formed by microstructure descriptors. Each DOE point represents one microstructure design, for which we reconstruct one or multiple statistically equivalent microstructures using a sequential descriptor-based microstructure reconstruction algorithm [14]. The properties of reconstructions are simulated using FEA. Metamodels, also known as surrogate models [53], are created to replace the computationally expensive FEA models in optimization. The optimal designs are verified by running simulations on the reconstructed microstructures.

We make two sets of comparisons to demonstrate that the performances of microstructure designs are improved by using key descriptors as microstructure design variables. (1) single-objective optimizations. Microstructure descriptors are used as design variables in single-objective optimizations to minimize \( L \), maximize \( P \), and maximize \( H \), respectively. We compare the single-objective optimal designs using the key descriptor set and the optimal designs using the empirical descriptor set in Fig. 8(\(a\)). Significant improvements in properties are achieved by using

![Fig. 7 Microstructure reconstructions using statistically learned descriptor set and empirical descriptor set](image)

![Fig. 8 Comparison of optimal designs (Min: \( L \), Max: \( P \), Max: \( H \)) using key descriptors and empirical descriptors. (a) Single objective optimization for each objective; (b) multi-objective optimization with equal weights on objectives. Two examples of optimal microstructures, Max \( H \) by key descriptors and Max \( H \) by empirical descriptors, are also plotted.](image)
statistically learned descriptors as design variables. Figure 8 also shows two examples of optimal microstructure designs, which are obtained using different design variable sets. The optimal microstructure design by the key descriptor set is more dispersed (have larger surface area) compared with the design by the empirical descriptor set. Larger surface area leads to a higher value of property \( h \) (2). Multi-objective optimization. Microstructure descriptors are used as design variables in multi-objective optimizations, which minimize \( L \), maximize \( P \), and maximize \( H \) simultaneously. Equal weights are assigned to the three normalized objectives. Better designs can be obtained by using key descriptors in multi-objective optimization as well, as shown in Fig. 8(b). It is observed that the optimal design using the key descriptor set has lower \( L \), roughly the same \( P \), and higher \( H \) compared with the optimal design using empirical descriptor set as design variables.

5 Conclusion

This paper presents a machine learning-based method for identifying key microstructure descriptors as microstructure design variables. It facilitates low dimensional descriptor-based microstructure representations. Starting from a complete set of microstructure descriptors collected from literature, we reduce the redundant descriptors via descriptor–descriptor correlation analysis and correlation function-based supervised learning. These two steps are computationally efficient as only image analyses are involved. Furthermore, a property-based supervised learning is conducted to identify the key descriptors. Microstructure design variables are chosen from the key descriptors based on their contributions to material properties as well as the need for minimizing descriptor–descriptor dependency. This four-step method enables parametric optimization of heterogeneous microstructures using a small set of physically meaningful descriptors to achieve target properties. We demonstrate this method using a case study of designing polymer nanocomposites’ microstructures. This proposed method leads to better designs compared with designs using descriptors chosen randomly or empirically.

Our research contributes to the computational design of microstructural materials system in the following three aspects. First, this method provides a rigorous way for material scientists to choose microstructure descriptors in analyzing and designing new materials. It is demonstrated that optimization using key descriptors obtained by machine learning improves the performance of microstructure designs for the interested type of polymer nanocomposites. The proposed method identifies descriptors that are important to both microstructure morphology and properties. Second, this method effectively cuts down the computation costs by introducing image analysis-based prescreening. The prescreening steps reduce the dimension of the candidate descriptor set before applying the property-based supervised learning. Conducting property-based supervised learning on the reduced descriptor set requires less number of samples (simulations) than on the full descriptor set. Third, this method enables parametric optimization of the microstructure with a small set of design variables. State-of-art computational design methods (e.g., DOE, metamodeling) are applied to explore the microstructural design space to achieve optimal material properties.

In future works, the proposed method will be further validated with different types of microstructural material systems. Another possible future work is to include the process–microstructure relation into the material design process to cover the whole spectrum of material design and to ensure that material engineers can fabricate the optimal microstructure. The machine learning approach will be applied to the process–microstructure database to establish mathematical relations between processing parameters and the resultant microstructures.

Appendix: Descriptor Parameters’ Sequence Numbers in Fig. 5

The tablebelow lists the number and name of the descriptors in Fig. 5. The three intracorrelated descriptor groups are highlighted using gray background color.

The number after each symbol represents the order of the moment (first: mean; second: variance; third: skewness; fourth: kurtosis).

### References


