Reduced-order structure-property linkages for polycrystalline microstructures based on 2-point statistics

Noah H. Paulson a, Matthew W. Priddy b, David L. McDowell a, c, Surya R. Kalidindi a, d, e

a George W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0405, USA
b School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0405, USA
c School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0245, USA
d School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA
e George W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

Abstract
Computationally efficient structure-property (S-P) linkages (i.e., reduced order models) are a necessary key ingredient in accelerating the rate of development and deployment of structural materials. This need represents a major challenge for polycrystalline materials, which exhibit rich heterogeneous microstructure at multiple structure/length scales, and exhibit a wide range of properties. In this study, a novel framework is described for extracting S-P linkages in polycrystalline microstructures that are obtained using 2-point spatial correlations (also called 2-point statistics) to quantify the material’s microstructure, and principal component analysis (PCA) to represent this information in a reduced dimensional space. Additionally, it is demonstrated that the use of generalized spherical harmonics (GSH) as a Fourier basis for functions defined on the orientation space leads to a compact and computationally efficient representation of the desired S-P linkages. In this study, these novel protocols are developed and demonstrated for elastic stiffness and yield strength predictions for α-Ti microstructures using a dataset produced through microscale finite element simulations.

Keywords:
2-Point correlations
Microstructure
Computational model
Crystal plasticity
Reduced-order model

1. Introduction
Materials development and deployment efforts require evaluation of the properties of large sets of potential material internal structures (generically referred to as microstructures; this description also includes information regarding chemical composition at the relevant length scales) [1–13]. Currently, experiments and simulations are used for such analyses at a high cost, severely limiting the extent of the materials design spaces explored [14,15]. In addition, most of the current approaches (e.g., finite element models) employed in modeling and simulation are not readily invertible, i.e., it isn’t easy to determine a candidate microstructure given a set of desired properties. Reduced order models are critical to the acceleration of materials discovery and development, as they enable high throughput exploration of the materials design spaces for rapid down selection [1,6,16–18]. In this work, we specifically focus on reduced order structure-property (S-P) linkages for polycrystalline microstructures that exhibit rich heterogeneity in the spatial distribution of the crystal lattice orientation at the mesoscale (also referred to as grain-scale). These microstructures exhibit strongly anisotropic elastic and inelastic properties at the macro-scale [5,16,19–31].

It is important to pursue the development of reduced order models within the context of established physically-based theories. There is indeed a rich literature on homogenization theories for polycrystalline microstructures that can be used to guide the development of surrogate models. For elastic stiffness or compliance components, the foundational concepts can be traced back to the elementary bounds established by Voigt [32], Reuss [33], and Hill [34], as well as the more sophisticated self-consistent approaches of Kroner [35], which leverage Eshelby’s solution [36]. For inelastic properties, analogous foundations can be traced to the works of Taylor [37] and Sachs [38]. Detailed discussion of these theories cast in a modern continuum mechanics framework can be found in Nemat-Nasser and Mori [39], Milton [40], Mura [41], Qu and Cherkaoiu [42] and Roters [43].

The central limitation of the theories mentioned above is that...
they can only incorporate a very limited amount of information regarding the material microstructure. To date, microstructural information included in the application of established composite theories to polycrystalline microstructures has largely been limited to the orientation distribution (also called texture) in the sample, and in some cases, averaged shape factors (based on idealizing the grains as ellipsoids) [31,44]. The main exception to these statements comes in the form of statistical continuum theories formulated originally by Brown [45] for electrical properties of materials, and subsequently introduced by Volkov and Klinshikh [46], Lomakin [47], and Beran and Molineux [48] for mechanical properties. These ideas have been further refined in later years by Beran [49,50], Kröner [51–53], Torquato [54–56], Adams [16,57,58], and others [59–63]. A unique feature of this formulation is that the effective property is expressed as a series expansion, where each term represents a contribution of a specific microstructural attribute expressed in the formalism of n-point statistics [57,58,64,65]. In some cases, the pre-factor for each term in the series can be evaluated using known Green’s functions [16,59]. Most importantly, this approach allows one to include as much detail of the material microstructure as one wishes in arriving at the homogenized properties of interest.

The sophisticated higher-order homogenization theories mentioned above have encountered some difficulties in their implementation. First, the Green’s functions required for the implementation of these theories are only available for a limited number of physical phenomena. Second, the convergence of the series is rather slow for high contrast composites [54,64]. Over the past decade, Kalidindi and coworkers [1,16,19,66–76] have developed a novel data science approach that still utilizes the basic mathematical formulation of the statistical continuum theories, and circumvents the central impediments mentioned above by employing modern data science tools. Mainly, it has been shown that the elusive pre-factors (related to Green’s functions) in the series expansion can be calibrated to results from numerical simulations (e.g., performed using finite element models). This new framework for establishing S-P linkages has been referred as material microstructure as one wishes in arriving at the homogenized properties of interest.

In order to construct S-P linkages, microstructural information must be presented in a form that allows for a rigorous quantitative comparison. Unfortunately, images of microstructure cannot fulfill this role in their raw representations. In other words, a pixel by pixel comparison of two random microstructures, either from the same sample or from different samples, will not reveal useful information. Traditionally, metrics such as grain size distributions and phase volume fractions have been used to capture the salient microstructural information [83,84]. However, these approaches do not consider the complexities of the microstructure geometry with sufficient fidelity for many properties of interest. In recent years, Kalidindi and co-workers have employed 2-point spatial correlations to evaluate microstructural distances [64], cluster microstructures [74,85], and to provide a basis for the construction of S-P linkages [68,70,73]. This higher order statistical representation of microstructure precludes the need to manually select microstructure features (and potentially introduce bias) in the construction of S-P linkages; the 2-point spatial correlations already include many of the conventionally defined metrics of the microstructure.

Following the framework developed by Adams et al. [5,16,57], the microstructure in hierarchical materials systems can be mathematically captured through the function \( m(h, \mathbf{x}) \), which reflects the probability density of finding local state \( h \) (within an incremental \( \Delta h \) around \( h \)) at the spatial location \( \mathbf{x} \). In this formalism, \( h \) and \( \mathbf{x} \) are both treated as continuous variables. However, in practice it is much more useful to employ a discretized description of the microstructure function. Moreover, most experimental characterization techniques recover only a discretized description of the microstructure (as a consequence of the resolution limits of the equipment). Recognizing this, Adams et al. [5,16,57] have proposed a discretized description of the microstructure function \( M^p_{\mathbf{x}} \) based on primitive binning of both the spatial domain and the local state space, i.e.,

\[
m(h, \mathbf{x}) \Delta h = \sum_{i=1}^{N} \sum_{s=1}^{S} M^p_{\mathbf{x}}(h) \chi_i(h) \chi_s(\mathbf{x})
\]  

(1) 

where the indicator function \( \chi_i(\cdot) \) is defined such that it is equal to one if and only if the argument belongs to the bin labeled \( i \) and zero otherwise, and \( \Delta h \) denotes the size of the local state bin employed. An important consequence of this definition is that \( M^p_{\mathbf{x}} \) has a very
simple interpretation - it is essentially the volume fraction of local states belonging to the bin \( n \) (in the local state space) present in the bin \( s \) (in the spatial domain). It is important to recognize that both \( n \) and \( s \) are simply indices that point to a specific local state and spatial bin (i.e., voxel), respectively.

Although \( M_{ps}^{n} \) has a simple interpretation and provides a useful description of the microstructure, it alone does not provide the salient information on the topological features of the microstructure. For this, one has to compute the spatial correlations. The most systematic approach to quantifying the spatial correlations in a given microstructure is through the use of the \( n \)-point spatial correlations framework \([54,60,65,73,78,86–88]\). The most basic spatial correlations can be computed in the form of 2-point statistics formally expressed as

\[
f(h, h'|r) = \frac{1}{\text{Vol}(\Omega_r)} \int_{\Omega_r} m(h, x)m(h', x + r)dx \quad (2)
\]

In Eq. (2), \( f(h, h'|r) \) denotes the 2-point statistics reflecting the conditional probability density associated with finding the local state \( h \) at the spatial location \( x \) and the local state \( h' \) at the spatial location \( x + r \). As before, \( f(h, h'|r)[\Delta h]^2 \) denotes the conditional probability. Note that the volume examined in evaluating the integral (denoted \( \Omega_r \)) is expected to depend on the choice of \( r \) (this is because of the need to know the local states at both spatial locations \( x \) and \( x + r \)). For the same reasons that were mentioned earlier, a much more practical form of the 2-point spatial correlations is a discretized version expressed as

\[
f(h, h'|\Delta h)^2 = \sum_{p=1}^{N} \sum_{n=1}^{N} \sum_{t=1}^{P^\text{np}} x_{hn}(h)x_{ht}(h')x_{tx}(r) \quad (3)
\]

Once again, the discretized version \( F_{tn}^{\text{np}} \) has a simple physical interpretation - it captures the conditional probability of finding ordered local states belonging to bins \( n \) and \( p \) (in the local state space) in spatial bins whose centers are separated by the set of vectors indexed by \( t \). Further details and a rigorous treatment of this framework can be found in Ref. \([57]\). The discretized forms of the microstructure function and the 2-point spatial correlations are indeed related by the following simple relation:

\[
F_{tn}^{\text{np}} = \frac{1}{|S_t|} \sum_{s=1}^{S_t} M_{ts}^{s} M_{ps}^{n} \quad (4)
\]

Specific attention needs to be afforded to several details of Eq. (4) before one endeavors to implement it in a computational algorithm: (i) the mapping of the vector \( r \) to the index \( t \) needs to follow specific protocols such as those outlined in Ref. \([57]\), (ii) the earlier discussion regarding \( \Omega_r \) translates to \( S_t \) to identify the set of valid trials involved in evaluating \( F_{tn}^{\text{np}} \), which depends on the vector index \( t \) \([89]\), (iii) it is often convenient to employ multidimensional index arrays for \( s \) and \( t \) to naturally index the spatial bins in 2-D and 3-D microstructures \([1]\), and (iv) \( |S_t| \) denotes the total number of spatial bins involved when \( s \) is used as a multidimensional array index.

For a complete representation of \( f(h, h'|r) \) from equation (3), \( F_{tn}^{\text{np}} \) must be calculated for all \( t \), \( n \), and \( p \). Depending on the size of the microstructure and choice of basis functions, this set can be large and unwieldy. For example, \( F_{tn}^{\text{np}} \) contains close to one-million features for a 3-dimensional microstructural sample with 21 elements per edge and 10 local state bins. At this stage, data-science techniques can be applied to reduce the dimensionality of the problem. One approach is to use principal component analysis (PCA), which is a distance-preserving linear transformation. PCA finds orthogonal and linear combinations of the original features such that they are ordered from highest to lowest contributors to the variance in the dataset \([77]\). Let \( j = 1, 2, \ldots, J \) index the individual elements (i.e., microstructure exemplars) in the dataset being studied. Let \( \{F_{jt}^{(i)} \mid t = 1, 2, \ldots, R \} \) denote a vectorized set of all spatial statistics of interest for a single microstructure exemplar. As an example, this set might include all of the 2-point statistics computed from Eq. (4) for all combinations of \( (t, n, p) \). This set can also include higher-order spatial statistics (e.g., 3-point statistics), if deemed essential for the specific study. \( R \) denotes the total number of features (i.e., microstructure measures) included in the analysis. With this notation, PCA allows an orthogonal decomposition of \( F_{jt}^{(i)} \) as

\[
F_{jt}^{(i)} = \sum_{\ell=1}^{\min[(J-1)/R]} \alpha_{j\ell} \varphi_{\ell} + F_t
\]

where \( \varphi_{\ell} \) are the PC basis vectors, \( \alpha_{j\ell} \) are the transformed coordinates of the exemplar in the space spanned by the PC basis vectors, and \( F_t \) is ensemble average of \( F_{jt}^{(i)} \). Note that the maximum number of principal components is limited by the smaller value of \( R \) and \( (J-1)/R \). However, in most practical applications, only a handful of \( \alpha_{j\ell} \) (called PC scores) are employed in the reduced-order representation of the original dataset. This representation may be written as \( \{\alpha_{j\ell} \mid \ell = 1, 2, \ldots, R \} \), where the truncation level \( R \) is objectively decided based on the variance explained from the original dataset. The low-dimensional PC scores of the microstructures have been successfully utilized in building searchable microstructure databases \([85]\) as well as in the construction of robust and reliable reduced-order S-P linkages \([68,70]\).

The introduction of PC representations of the n-point statistics in the series expressions developed in statistical continuum theories for effective properties produces a surprisingly simple expression \([1,68]\), i.e.,

\[
P_{\text{eff}} = A_0 + \sum_{i=1}^{R} A_i \alpha_i
\]

where \( A_i \) capture the underlying physics of the problem, \( P_{\text{eff}} \) is the effective property of interest, and \( R \) is the truncation level selected for the PC representations. Indeed, Eq. (6) expresses a linear relationship between \( P_{\text{eff}} \) and the scores of a chosen set of PC basis vectors (which themselves are extremely rich sets of microstructural features). Although the above equation is easily derived for elastic response of composite material systems, it can be extended to nonlinear effective properties with suitable modifications \([1,40,59,60,68,70,73,74]\). In the MKS framework, the \( A_i \) coefficients are established through regression using datasets generated by numerical simulation tools (e.g., finite element models) on selected microstructure exemplars.

### 3. Extensions to the theoretical framework

Local states for the general class of polycrystalline metals may include descriptors such as crystal symmetry, dislocation density, and chemical composition (among others). However, for single-phase annealed polycrystals, the main local state of interest at the grain-scale is the crystal lattice orientation, \( g \). The corresponding local state space is simply the orientation space. Based on the concepts discussed thus far, a simple approach to address polycrystalline microstructures is to bin the fundamental zone of orientation space \([57]\). As a specific example, the fundamental zone of orientation space for hexagonal crystals is expressed as
relations highly unwieldy. Computation, storage, and utilization of the 2-point spatial correlations in polycrystalline microstructures. In this approach, we retain the primitive binning for the spatial variables and only change the basis for the orientation variables. With these changes, the microstructure function can be expressed as

\[ M_{is}^{mv} = \sum_{l,m,r} \sum_{s} M_{is}^{mv}(g) \chi_{s}(x) \chi_{r}(x) \chi_{l}(g) \chi_{m}(g) \]  

where \( \chi_{s}(x) \) denote symmetrized GSH functions (in the present work, the symmetry of interest is the hexagonal-triclinic symmetry, where the first symmetry refers to crystal symmetry and the second one to the sample symmetry), * denotes a complex conjugate, \( |\Delta x| \) is the uniform size (volume) of a spatial bin, and \( M_{is}^{mv} \) are the GSH Fourier coefficients. As a special case, when there is a single crystal of lattice orientation \( g_0 \) in a spatial bin \( s \), the corresponding GSH Fourier coefficients are simply given by

\[ M_{is}^{mv} = \begin{cases} 2(1 + \frac{1}{|\Delta x|}) & \text{if } s = s_0 \\ 0 & \text{else} \end{cases} \]

For simplicity of notation, we will map every distinct combination of \((l, m, r)\) to a single index \( l \) in all of the ensuing equations. As a result of this simplification, \( M_{is}^{mv} \) will be henceforth denoted simply as \( M_l \). Extending the concept above, the description of the 2-point spatial correlations of orientations in the polycrystalline microstructure is expressed as

\[ f(g, g') = \frac{1}{S_i} \sum_{S_k} M_{k}^{LK} \left( \sum_{l} T_{l}(g) T_{l}(g') \chi_{l}(x) \chi_{l}(x') \right) \]  

However, if one were to start with the formal definition of the 2-point spatial correlation (Eq. (2)) and substitute the GSH representations of the microstructure function from Eq. (8), we can derive the following relationship between the GSH Fourier coefficients in Eqs. (8) and (10):

\[ P_{l}^{LK} = \frac{1}{S_i} \frac{S_k}{S_{s-1}} M_{k}^{LK} M_{s-1}^{L} \]

The implicit convolution in Eq. (11) (same as in Eq. (4)) allows the exploitation of the FFT algorithms in computing the 2-point spatial correlations [64,85–87]. As mentioned earlier, the set of 2-point statistics computed using Eq. (11) will be large and unwieldy. As before, we will seek a lower dimensional representation of this set using PCA. In this case, however, the Fourier coefficients computed using Eq. (11) are complex. One can either perform the PCA directly on these complex representations, or reshape them into separate real and imaginary components prior to performing the PCA. Both of these approaches would be completely equivalent, especially given that half of the \( P_{l}^{LK} \) coefficients are complex conjugates of the other half in our application. In order to construct the S-P linkage, we would continue to use the model form shown earlier in Eq. (6) and calibrate the physically-based coefficients \( A_i \) with numerical simulation results obtained from finite element models for a selected number of microstructure exemplars.

4. Case study

The purpose of this case study is to demonstrate the viability and potential of the data-driven protocols developed in the previous sections for the creation of S-P linkages. In these approaches, the physics capturing coefficients of Eq. (6) will be determined through the following steps: (i) An ensemble of microstructure exemplars will be generated which exhibit a rich and diverse set of topological features (these might include grain size, shape, and orientation distributions) covering a range of microstructures of interest for the intended application of the S-P linkages being produced. (ii) Crystal plasticity finite element (CPFE) simulations are performed on each microstructure exemplar in the ensemble to extract the effective (bulk) properties of interest. (iii) Each microstructure exemplar is quantified using 2-point spatial statistics with GSH basis functions. (iv) PCA is performed on the 2-pt statistics representations of the microstructure exemplars in the ensemble to produce a low-dimensional representation of each exemplar microstructure in the form of their PC scores. (v) Multivariate linear regression is performed on the assembled dataset (from steps (i)-(iv)) to calibrate the physics capturing coefficients of Eq. (6). This calibrated model may then be validated by comparing the CPFE computed properties for new microstructure exemplars against the predictions of the reduced-order model. Each of the steps described above will be detailed in the following sections.

4.1. Generation of microstructure ensemble

In this work synthetic microstructure volume elements (MVEs) are generated both for establishing the S-P linkages of interest and for validating them. The MVEs must be sufficiently large to capture the rich microscale interactions between grains of different sizes, shapes, and orientations. In prior work, it was shown that these microscale interactions are most efficiently captured through influence kernels [19,66,67,69,71,72,75,76]. More specifically, the decay in these Green’s function-based influence kernels provides guidance for the minimum size of the MVEs. It is generally noted that the size of the MVEs should be larger for material systems with higher contrast in the local properties of the microscale constituents. Based on our prior experience [19,66,69], we employ MVEs of size \( 21 \times 21 \times 21 \) voxels in this work, where each voxel has a side length of \( 10 \mu m \).

The MVEs used in this work are generated using the open-source DREAM.3D package [92], which accepts as inputs microstructure measures such as grain size, grain shape, orientation and misorientation distributions. In this first study on polycrystalline microstructures, we will focus our attention on the distribution of grain orientations (this is the main reason for using the GSH representations described earlier). Consequently, we restrict our attention to MVEs with different textures, but exhibiting a common collection of experimental characterizations of polycrystalline titanium materials accessible to us. In future work, it would be desirable to expand the range of the generated MVEs using more sophisticated approaches [93].

The targeted grain size distribution for the synthetically
generated MVEs follows a log-normal distribution with an average grain size of $30\mu m$ and a $15\mu m$ standard deviation. Each MVE is comprised of an average of 215 distinct grains. Seven unique textures are selected as targets for the synthetic microstructure generation. These textures are inspired by those seen in our prior work [94,95]. The target textures are entered into DREAM.3D as lists of orientations expressed as triplets of Euler angles with associated weights and spreads. DREAM.3D is then utilized to randomly generate thirty polycrystalline MVEs for each microstructure class through an ellipsoidal packing algorithm. Then, DREAM.3D matches the target texture information for each microstructure class. Fig. 1 shows example MVEs and (0002) pole figures for each microstructure class (corresponding to each selected texture). This figure only enables a qualitative comparison of microstructures. However, in later sections, the 2-pt spatial correlations and PCA will be used to quantify and visualize the differences between these microstructures.

4.2. CPFE simulations

In this work, CPFE simulations are used to extract the homogenized properties for each MVE. The combination of each MVE and its associated properties constitutes one data point for the desired S-P linkages. The crystal plasticity framework used here was developed in prior work for S-P linkages. The crystal plasticity framework used here was carried out such a way that the complete set of terms corresponds to any selected value of $I$ are included as a group. Other words, if one truncates at $I = 4$, then it is important to include all $I$ terms corresponding to all combinations of $\mu$ and $\nu$ for $I \leq 4$ [1,16]. For the hexagonal-triclinic GSH basis employed in this work, adaption of this consideration results in acceptable truncation levels of $L = 1$, $15$, $41$, and $90$ (these correspond to $I$ values truncated at $2$, $4$, $6$ and $8$, respectively).

Once a truncation level has been selected for $L$, the Fourier coefficients of the 2-point statistics may be computed using Eq. (11). Let us denote this truncation value as $L$. In employing Eq. (11) one can take advantage of a few strategies known to dramatically reduce the computational cost involved. First, the number of 2-point spatial correlations is $L^2$ (this includes all possible pairings of indices $L$ and $K$ in Eq. (11)). In prior work on the multiphase microstructures [87], analytic relationships were derived between the spatial correlations. Appendix A presents these interrelationships for the 2-point statistics computed here on polycrystalline microstructures, where it is shown that the number of independent correlations is $2L - 1$. It is important to recognize that these dependencies are not all linear. Furthermore, since PCA automatically removes the linearly dependent correlations, we do not need to worry about the linearly dependent correlations. By repeated trials, the optimal set of spatial correlations to represent each MVE was selected (further details are provided in Appendix A). It was determined through these trials that the specific set of $2L - 1$ correlations identified in Appendix A produces the best compromise between accuracy and computational cost for the S-P linkages produced in this work.

4.4. Reduced dimensional microstructure representation

Once spatial statistics are computed for all MVEs in the selected ensemble, PCA is performed to obtain reduced dimensional representations of the microstructure statistics. In this study, there are far more features (coefficients in the Fourier representation of the spatial statistics) than samples (MVEs). Therefore, the maximum number of PC scores that can be computed is 209 (one less than the number of MVEs used to build the PC space; see Eq. (5)) [74]. However, a much smaller number of PCs are required to satisfactorily represent the spatial statistics for this case study. Fig. 2 shows the cumulative variance captured by the inclusion of additional PCs. As $R$ increases, the cumulative explained variance (representing the 2-point statistics information) asymptotically approaches 100%. In addition, as the number of GSH basis functions used in the Fourier representation of the 2-point statistics increases (improving the fidelity of the microstructure representation), so does the $R$ required to achieve a given explained variance.

PCA also serves as an effective method to cluster MVEs and to
evaluate the distances between them \cite{68,70,73,74,85}. It is important to note that PCA results in unsupervised classification as the MVEs are not identified or labeled in any manner. Fig. 3a demonstrates this natural clustering of the MVEs in the first two PC dimensions (the dimensions with the largest and the second-largest variances) for GSH spatial statistics at a truncation level of $L = 15$. Remarkably, even in just two dimensions, the MVEs distinctively cluster by their associated microstructure class (here differentiated exclusively by target texture). In other words, even though the PCA is not informed explicitly about the textures in the MVEs, the first two PC scores effectively cluster the MVEs based on their differences in texture (note that the texture information is included in the 2-point statistics). Furthermore, it is seen that the intraclass variance in each cluster is not the same. For example, the MVEs belonging to class A exhibit more differences among themselves compared to the MVEs of class G in the first two PCs. This is, of course, a consequence of how each MVE is randomly generated for target microstructure statistics in the DREAM.3D software. Comparing Figs. 1 and 3a, it is also seen that microstructures with similar textures are nearby in the PC space, while microstructures with dramatically different textures are well separated. Notice that microstructure classes with strong basal texture components are located towards the left of Fig. 3a, while microstructure classes with strong transverse texture components are located towards the right of this figure. Although such observations provide intuition into the meanings of the different PCs, it should be noted that the complete information captured by each PC is actually stored in the vectors, $\mathbf{v}_i$, obtained in PCA (see Eq. (5)). Given the large size of these vectors (for a truncation level of $L = 15$ each PC has a length of 268,569 in the present study), it is not straightforward to establish their precise physical meaning.

Fig. 3b shows the relative distances between the centers of the microstructure clusters in PC space, presented by way of a

### Table 1

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c_{11}$</td>
<td>172.8 GPa</td>
<td>$\rho_{\text{basal}}$</td>
<td>350 MPa</td>
</tr>
<tr>
<td>$c_{12}$</td>
<td>97.91 GPa</td>
<td>$\rho_{\text{prism}}$</td>
<td>275 MPa</td>
</tr>
<tr>
<td>$c_{13}$</td>
<td>73.43 GPa</td>
<td>$\rho_{\text{pyramidal}}$</td>
<td>470 MPa</td>
</tr>
<tr>
<td>$c_{33}$</td>
<td>192.3 GPa</td>
<td>$\rho_{\text{pyramidal}} + c_{44}$</td>
<td>570 MPa</td>
</tr>
<tr>
<td>$c_{44}$</td>
<td>49.70 GPa</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 1. Sample MVEs (displayed using Grain ID maps) and (0002) pole figures for the generated microstructures classes A through G.

Fig. 2. Percentage PCA explained structural variance versus $R$ and $L$. 

---

dendrogram. The advantage of this representation is that the displayed distances are computed over all available dimensions (in the present work this is 209). From the dendrogram, it can be seen that the first two PC dimensions adequately capture the relative distances between the microstructure classes. This is indeed the power of PC representations; they provide high value low dimensional representations of practical worth in quantifying the microstructure statistics. As such, this is an important component in the effort to extract robust and reliable S-P linkages.

4.5. Extraction of structure-property linkages

The low dimensional representation of each MVE and its associated properties established using the CPFEmodel described earlier constitute the calibration dataset for extracting the desired S-P linkages. More specifically, we intend to calibrate the physics-capturing coefficients of Eq. (6) using a simple multivariate linear regression. Prior to calibration, two critical hyper-parameters, \(\hat{L}\) and \(\hat{R}\) must be selected. These hyper-parameters cannot be arbitrarily chosen; they must be selected individually for each property of interest by systematically exploring a large number of potential values (for both parameters) and evaluating their efficacy in producing the most reliable S-P linkages.

In order to accomplish this step, it is necessary to select the metrics by which we objectively evaluate each potential S-P linkage produced in this work. One obvious metric, called the calibration error, is simply defined as the mean of the absolute error between the MKS predicted and the CPFEM simulated values of the property for all MVEs, normalized by the mean value of the simulated property. Fig. 4 shows the calibration error for \(\hat{R}\) up to 60 and four different \(\hat{L}\) levels for elastic stiffness (Fig. 4a) and yield strength (Fig. 4b). It is clear from these plots that the gains in accuracy quickly diminish with increasing values of both \(\hat{R}\) and \(\hat{L}\). This observation attests to the power of the PC and GSH representations employed in this work. Note also that the average error values are remarkably low - only about 0.2% for the elastic stiffness prediction and about 1.5% for the yield strength prediction. The fact that the prediction error for the yield strength is higher than that for the elastic stiffness is quite reasonable and expected. Indeed, prior work has shown that the elastic response of polycrystalline microstructures can be represented accurately by a limited set of GSH basis functions [29,69]. In the present work, we observe no significant improvement in the accuracy of the elastic S-P linkage beyond \(\hat{L} = 15\) (corresponding to \(\hat{L} \leq 4\)). For the predictions of the yield strength, higher values of \(\hat{L}\) did provide a modest improvement in the accuracy. These observations of the efficacy of the GSH representations are highly consistent with prior related work [29,69,80].

Considering the effect of the value of \(\hat{R}\), the most dramatic decrease in error occurs within the inclusion of the first 10 PCs. This is consistent with the observation in Fig. 2 that the explained structural variance quickly approaches 100% with increasing numbers of PCs.

Fig. 4a and b illustrate the potential for overfitting the S-P linkages. The calibration error is usually not a good indicator for overfitting as it always decreases for increasing values of hyper-parameters such as \(\hat{R}\). In prior work, the Leave-One-Out Cross Validation (LOOCV) error metric was employed as an indicator of overfitting [68,70]. In this approach, the S-P linkage is calibrated using the data for all MVEs in the ensemble except for one, and the error in the prediction of the property for the excluded MVE is recorded. The entire procedure is repeated for each MVE in the ensemble. The LOOCV error is computed as the mean of the prediction errors for each excluded MVE in the ensemble and normalized the same way as the calibration error. The normalized LOOCV error is plotted for elastic stiffness and yield strength linkages in Fig. 4c and d, respectively. For both properties, the LOOCV error drops quickly for small values of \(\hat{R}\). The LOOCV error, however, begins to increase at higher values of \(\hat{R}\) indicating overfitting. It is also observed that the minimum LOOCV error occurs at a higher value of \(\hat{R}\) for the higher GSH truncation levels. This is once again consistent with the observation in Fig. 2 that a higher \(\hat{R}\) is required to achieve the same explained variance for at higher values of \(\hat{L}\).

The optimal selection of \(\hat{L}\) and \(\hat{R}\) depends on the desired characteristics of the S-P linkages to be produced. If the minimization of error is paramount, the linkage with the lowest LOOCV error should be chosen. Such a linkage is expected to perform the best for new MVEs. However, if compactness is desired, then the linkage which has the minimum \(\hat{L}\) and \(\hat{R}\) yet has a LOOCV error below some acceptable threshold should be selected. In this study, reduced-order (low-dimensional) linkages are prioritized to demonstrate their applicability for inverse materials design problems. Through the consideration of Fig. 4c a linkage for elastic stiffness is selected with 3 PCs and 15 GSH basis functions. From Fig. 4d a linkage for yield strength is selected with 6 PCs and 41 GSH basis functions. This means that the S-P linkages produced in this study for elastic stiffness and yield strength only have 4 and 7 coefficients, respectively, in Eq. (6).

4.6. Validation of structure-property linkages

In the previous sections, a novel data science framework is developed and implemented for the extraction of highly accurate
and robust S-P linkages (note the very low error values in Fig. 4). We now seek to further validate these linkages using completely new MVEs (those that have not been used in the calibration). For this purpose, a new set of MVEs are generated and their properties are simulated using the same protocols described earlier in this paper. However, in an effort to make these MVEs different from the calibration MVEs, five additional target textures are employed. Example MVEs and (0002) pole figures for each of the new microstructure classes are presented in Fig. 5a. Microstructure class H is inspired from literature [95], while classes I through L are created using different combinations of texture components present in classes A through H. The new microstructure classes are intended to be interpolations between existing classes. The S-P linkages may give good results for small extrapolations beyond the bounds of the calibration MVE ensemble, but cannot be expected to predict the properties of completely new microstructures. As before, thirty MVEs are generated for each of the five new microstructure classes. Furthermore, thirty additional MVEs are generated for each of the existing microstructure classes. CPFEM simulations are performed for all MVEs in the validation ensemble to extract their yield strength and elastic stiffness.

For each MVE in the validation ensemble, the 2-point spatial correlations are computed up to the GSH truncation levels selected for each property in Sec. 4.5. The spatial statistics for each new MVE are then transformed into the reduced dimensionality PC space already established in Sec. 4.4. This validation ensemble is displayed in Fig. 5b in the PC1-PC2 space. For better visualization, the MVEs from microstructure classes H through L are highlighted in color in this figure, while MVEs from the original microstructure classes are shown in gray. The properties of each MVE in the

![Image](image-url)

**Fig. 4.** Error in the property prediction versus number of PCs included in the S-P linkage for different selections of the number of GSH basis functions. (a) Error in the prediction of elastic stiffness for MVEs in calibration set. (b) Error in prediction of yield strength for MVEs in calibration set. (c) LOOCV error in prediction of elastic stiffness for MVEs in calibration set. (d) LOOCV error in prediction of yield strength for MVEs in calibration set.

![Image](image-url)

**Fig. 5.** (a) Sample MVEs (with Grain ID displayed) and (0002) pole figures for microstructure classes H through L. (b) The ensemble of validation MVEs in PC1 and PC2 (for \( L = 15 \)). MVEs from previously existing microstructure classes are colored gray.
validation ensemble are then predicted using the calibrated S-P linkages obtained in Sec. 4.5. Fig. 6 presents parity plots for each S-P linkage, where the predicted property is plotted versus the simulated property (via CPFE simulations) for each MVE in the validation ensemble. The maximum prediction errors for elastic stiffness and yield strength are highly consistent with each other, and are below 1% and 5%, respectively. The success of both linkages in accurately predicting properties for these new microstructure classes demonstrates the robustness of the S-P linkage produced in this work.

5. Conclusions

In this work, novel data science protocols for the construction of reduced-order structure-property linkages in polycrystalline materials were developed and validated for the low crystal symmetry α-titanium materials system. This is the first use of GSH basis functions in the MKS framework for the computation of spatial statistics and for the construction of homogenization linkages. Protocols to predict both elastic stiffness and yield strength were developed, each in the form of polynomial equations with a small number of coefficients; it was demonstrated that only four and seven coefficients were required to predict elastic stiffness and yield strength respectively. These linkages provided massive computational savings versus traditional protocols. CPFE simulations for the prediction of yield strength in a single MVE took two hours on four processors on a super-computing cluster, while the reduced-order S-P linkage required only three seconds with one processor. This represents nearly four orders of magnitude reduction in processing time. The simplicity and efficiency of these robust structure-property linkages are well suited for inverse design protocols and will be instrumental in future materials design efforts.

Acknowledgments

This work was supported by the National Science Foundation under Grant No. CMMI-1333083. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. NHP would like to thank Dipen Patel and David Brough for many illuminating conversations throughout the development of this work.

Appendix

In previous work, Niezgoda et al. [87] identified redundancies in the full set of 2-point spatial statistics. Knowledge of these redundancies is crucial for identifying an efficient yet comprehensive set of spatial statistics for the representation of microstructure. Specifically, it was shown that the discrete Fourier transforms (DFTs) of the 2-point statistics, \( F_{j}^{np} = \mathcal{F}(p^{np}) \), exhibit the following interrelationships:

\[
\begin{align*}
F_{j}^{np} F_{j}^{pq} & = F_{j}^{np} F_{j}^{nq} , \\
F_{j}^{np} & = F_{j}^{nq} , \\
F_{j}^{nq} & = F_{j}^{p} F_{j}^{p} F_{j}^{q} \\
& = F_{j}^{p} F_{j}^{q}
\end{align*}
\]

where * denotes the complex conjugate. It is convenient to visualize \( F_{j}^{np} \) in the array \( F_{j}^{np} \) where \( N \) is the number of discrete local states (in conventional binning of the local state space using indicator functions), Eq. (12) implies that if \( F_{j}^{np} \) is known for a row or column of the correlation array, then the remainder of the array may be calculated. Furthermore, due to the properties of the indicator basis in the representation of the local state, only \( N - 1 \) of the correlations in a given row or column are independent. Therefore there are only \( N - 1 \) independent correlations in the set of the 2 point spatial correlations for an N-phase microstructure.

When the GSH basis is selected to describe the functional dependence of the grain orientation in the microstructure function, \( F_{j}^{PL} \) (see Eq. (11)) does not have a simple interpretation as a probability. As a result, the conclusions of Eq. (12) need to be suitably modified. Recognizing that \( T_{i}(g) = 1 \) it can be shown that

\[
\begin{align*}
\tilde{F}_{j-0}^{Q1} & = \tilde{F}_{j-0}^{TQ} = C_{Q} , \\
\tilde{F}_{j-0}^{Q1} & = |S| , \\
\tilde{F}_{j-0}^{Q1} & = \tilde{F}_{j-0}^{TQ} = 0
\end{align*}
\]

where \( C_{Q} \) are related to the 1-point statistics of the microstructure for single-phase polycrystals and contain the same information as the ODF. At this point it is useful to introduce two different arrays of GSH-based 2-pt correlations:

\[
A = \left[ \tilde{F}_{j}^{K} \right]_{L-1} \quad B = \left[ \tilde{F}_{j}^{LK} \right]_{L-1} \quad \text{with } L = 1 \text{ and } K = 1 \text{ excluded}
\]

Fig. 6. Predicted versus simulated response for (a) elastic stiffness with 3 PCs and (b) yield strength with 6 PCs. Grey markers are for validation MVEs generated with target microstructure statistics from the original seven microstructure classes.
Equations (12) and (13) imply that the independent set of correlations consists of the first row or column of $A$ in addition to any row or column of $B$ (for $2L - 1$ total correlations).

To select the optimal set of correlations for this study, the errors associated with the S-P linkages calibrated using different sets of correlations are compared. These test linkages are developed for both elastic stiffness and yield strength at a GSH truncation level of $L = 6$. The following list describes a selection of correlation sets and the quality of the resultant linkages.

1. All correlations in A ($L^2$ correlations): This results in the lowest error of all sets.
2. The lower triangular correlations of $A$ ($L^2/2L$ correlations): The error response is similar to set 1.
3. The first columns of $A$ and $B$ and the diagonal elements of $B$ ($3L - 1$ correlations): This results in an error slightly higher than sets 1 and 2.
4. The first column and diagonal elements of $A$ ($2L - 1$ correlations): This results in an error response similar to set 3.
5. The first columns of $A$ and $B$ ($2L - 1$ correlations): This results in an error higher than sets 1 and 2 but slightly better than 3 and 4.
6. The first column of $A$ ($L$ correlations): This results in higher errors than in the preceding sets. Furthermore, error does not decrease beyond the inclusion of the first several PCs. This indicates that the ODF information is important, but not sufficient for high-quality predictive capability.
7. The first column of $B$ ($L - 1$ correlations): This results in extremely high errors (the linkages had no predictive utility). It is hypothesized that the correlations in $B$ do not contain any information from the ODF, which is of first-order importance.

As a result of this investigation the first columns of $A$ and $B$ are selected as a good compromise between computational efficiency and S-P linkage quality.

References
