Available online at www.sciencedirect.com

ScienceDirect

Acta Materialia 91 (2015) 239–254

Acta MATERIALIA

www.elsevier.com/locate/actamat

Structure–property linkages using a data science approach: Application to a non-metallic inclusion/steel composite system

Akash Gupta,[a,1] Ahmet Cecen,[b,1] Sharad Goyal,[a,2] Amarendra K. Singh[a] and Surya R. Kalidindi[b,c,*]

[a] *TRDDC-TCS Innovation Labs, Tata Consultancy Services Ltd., 54 B, Hadapsar Industrial Estate, Pune, Maharashtra 411013, India*
[b] *School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA*
[c] *Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA*

Received 17 January 2014; revised 20 February 2015; accepted 22 February 2015
Available online 31 March 2015

Abstract—Practical multiscale materials design is contingent on the availability of robust and reliable reduced-order linkages (i.e., surrogate models) between the material internal structure and its associated macroscale properties of interest. Traditional approaches for establishing such linkages have relied largely on computationally expensive numerical simulation tools (e.g., the finite element models). This work investigates the viability of establishing low (computational) cost, data-driven, surrogate models for previously established numerical multiscale material models. This new approach comprises the following main steps: (1) generating a calibration (i.e., training) dataset using an ensemble of representative microstructures and obtaining their mechanical responses using established physics-based simulation tools (e.g., finite element models), (2) establishing objective, reduced-order, measures of the microstructures (e.g., using *n*-point spatial correlations and Principal Component Analysis), and (3) extracting and validating sufficiently accurate, computationally low-cost, relationships between the selected microstructure measures and effective (homogenized) properties (or performance metrics) of interest using various regression methods. In this paper, the viability of the data science approach in capturing such linkages (expressed as metamodels or surrogate models) for inelastic effective properties of composite materials is demonstrated for the first time.

Keywords: Homogenization theories; *n*-Point statistics; Data science; Microstructure measures; Principal Component Analysis

1. Introduction

Homogenization theories [1–15] are generally used to extract quantitative relationships between the effective properties of a composite material and certain important measures of its hierarchical internal structure. However, the accuracy of these theories is largely dependent on the adequacy of the selected set of microstructure measures. For example, it is well known that when a material's internal structure is quantified using only the volume fractions of various constituent phases, one can rigorously predict only the bounds of certain effective properties of the material. In general, one might expect to obtain better bounds and estimates of the effective properties, if one uses more information on the microstructure. Nevertheless, the results of the homogenization theories, when expressed as computationally efficient structure–property linkages, serve a critical role in hierarchical multiscale materials models (e.g., [16–18]).

In practice, rigorously formulated homogenization theories are currently available only for a fairly limited class of effective properties [1,8,19,20]. In most situations, one is forced to resort to numerical approaches (e.g., finite element models) for exploring the complex linkages between a material microstructure and its associated effective properties. When homogenization is accomplished through sophisticated (physics-based) numerical approaches, there is a critical need to distill the implied structure–property linkages from such models into low cost metamodels that can be easily incorporated into hierarchical multiscale modeling strategies. Currently, there is no established formalism for addressing this critical need.

The central challenge in the extraction of structure–property linkages from sophisticated numerical models is the identification of the salient microstructure measures that have a dominant influence on the macroscale properties of interest. Generally, this selection is made based on prior knowledge and intuition. For example, in establishing structure–property linkages for the effects of inclusions on the properties of steel (this case study will be explored in detail in this paper), it is typical to employ microstructure measures such as volume fraction, average inclusion size, and

* Corresponding author at: Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA. Tel.: +1 404 385 2886; e-mail: surya.kalidindi@me.gatech.edu
[1] Both authors contributed equally to this study.
[2] Currently working at Research and Development Centre, CEAT Ltd.

http://dx.doi.org/10.1016/j.actamat.2015.02.045
1359-6462/

average inclusion spacing (mean free path) as the salient microstructure measures [21–25]. However, one can easily see that this simple set of microstructure measures is unlikely to be the best possible set or even an adequate set, because it is easy to imagine multiple instantiations of microstructures that would exhibit the same values of these simple microstructure measures while displaying vastly different values of the macroscale properties of interest. This is particularly true when establishing structure–property linkages for defect-sensitive, potentially anisotropic, macroscale properties. In this paper, we explore the viability of employing a novel data science approach for objective (data-driven) formulation of such structure–property linkages.

In recent years, Kalidindi and co-workers [1,20,26–30] have presented a rigorous mathematical framework for systematic quantification of microstructure based on the concepts of n-point spatial correlations (also called n-point statistics) [6,31–34]. A major benefit of this approach is that it provides a naturally organized set of measures of the microstructure. However, the complete set of n-point statistics is an extremely large, unwieldy set, with exploding information content as a function of n. Even for $n = 2$, the number of spatial statistics is typically large enough that a rigorous analysis of large microstructure datasets and computationally efficient mining of structure–property linkages is only possible with the application of data analytic tools that include various combinations of feature selection and/or dimensionality reduction algorithms. One such tool that was previously employed with success [28–30,35–39] is the Principal Component Analysis (PCA) method. It is emphasized here that although PCA dimensionality reduction techniques have been explored in materials problems in prior literature [35–38], they have only recently been employed on 2-point spatial correlations of microstructure in attempts to successfully extract high fidelity structure–property linkages. The prior case studies have demonstrated the applicability of this new approach for (i) effective diffusivity of the porous transport layers in a Polymer Electrolyte Fuel Cell (PEFC) [40], and (ii) the effective elastic stiffness components of a porous elastic solid [28]. Building on these advances, we explore the applicability of these novel approaches for the first time to nonlinear phenomenon such as plasticity in a non-metallic inclusion/steel composite material system. Readers can refer to Appendix A for quick reference to list of important symbols and their definitions used in this paper.

2. Conventional structure–property linkages in inclusions/steel composites

Inclusions are defects (mainly oxides, sulfides, and nitrides) which are spatially distributed in the steel microstructure and are known to adversely affect various macroscale mechanical properties both during processing as well as in the final finished product [41–51]. Performance and properties of a steel sheet are strongly influenced by the size, shape, composition, type (hard or soft) and distribution of non-metallic inclusions in the steel sheet. Prior studies [21–25] aimed at extracting structure–property linkages for non-metallic inclusions/steel composites have largely relied on experiments, aided by analytical and numerical models. More importantly, all of the prior linkages have utilized highly simplified measures of microstructure such as volume fraction of inclusions, average inclusion size, and average inclusion spacing in establishing the desired linkages. For example, experiments conducted by Garrison and Wojcieszynski [22] have suggested that the toughness of the inclusion/steel composite is proportional to $f^{-\frac{1}{3}}$ (where f is the inclusion volume fraction) or R^{-1} (where R is the average inclusion radius). Experiments by Lankford [24] have suggested that the fatigue life of the inclusion/steel composite is reduced by a factor proportional to $D^{1/3}$ (where D is the inclusion diameter), provided the inclusions do not touch the surface and exceed a critical size. Murakami [25] treated inclusion as mechanically equivalent to small cracks and small defects, and has suggested a power-law relationship between fatigue strength and inclusion size. Gupta et al. [23] used numerical simulations to extract a modified material model for the inclusion/steel composite as:

$$\sigma_f = \sigma_S(1 - V_i)^{2.88} \tag{1}$$

where σ_f is the equivalent stress in steel matrix with inclusion, σ_s is the equivalent stress in the steel matrix, and V_i is the inclusion volume fraction. Baker et al. [21] proposed that short transverse toughness expressed as crack opening displacement COD is proportional to inclusion separation distance, d, in direction of fracture.

The above discussion points to the main limitations of the current approaches in capturing the structure–property linkages in the inclusion/steel composites: (i) all of the prior studies have used very simple isotropic (averaged over all directions) measures of the microstructure that reflect only averaged values (for example using average inclusion size instead of the inclusion size distribution). (ii) As a direct consequence of the use of the isotropic structure measures, the predicted macroscale properties are implicitly assumed to be isotropic. (iii) Most of the prior studies have considered only one structure measure at a time. Since the different structure measures (e.g., volume fraction, inclusion size, inclusion spacing) are not completely independent of each other, it is very difficult to aggregate all of this legacy information and extract structure–property linkages that utilize a common, extensible, set of structure measures. The data science approach described in this paper addresses these key limitations.

3. Data-science approach for establishing structure–property linkages

In recent work by Kalidindi and co-workers, a novel framework has been formulated and explored for establishing computationally low-cost structure–property linkages for a wide variety of hierarchical material systems exhibiting varied phenomenon. Our approach is grounded in the statistical continuum theories originally proposed by Kroner [52] and subsequently refined by Milton [8], Beran and co-workers [53,54], and Adams and co-workers [1,2,20,55–58]. In these rigorously formulated microstructure-sensitive homogenization theories, the effective properties take on the following characteristic series expression:

$$C^* = \overline{C} - \overline{C'\Gamma C'} + \overline{C'\Gamma C'\Gamma C'} - \ldots \tag{2}$$

where C^* denotes the effective property (defined at the higher length scale) and each term in the infinite series (the specific formulation in Eq. (2) being generally referred to as the weak-contrast expansion [1]) represents a highly

complex convolution integral over the microstructure volume. As an example, the second term in the series is defined as:

$$\overline{C'\Gamma C'} = \frac{1}{vol(\Omega)} \int_{\Omega} \int_{\Omega} C'(x)\Gamma(x,x')C'(x')dx'dx \tag{3}$$

where $C'(x)$ denotes the perturbation in the local property (compared to a selected reference property) at spatial position x in the microstructure volume Ω and $\Gamma(x,x')$ is a symmetric second derivative of a suitably defined Green's function for the underlying physics of the problem at hand. Although Eq. (2) was originally derived for the elastic response of composites, it has been subsequently shown to be also applicable to nonlinear material response [8,55,56] with appropriate re-interpretation of the terms in the series.

The properties of the Green's function help us recognize that $\Gamma(x,x')$ can be expressed as $\Gamma(x-x')$ or $\Gamma(-t)$, where $t = x' - x$. Employing a statistical representation of the microstructure based on the n-point spatial correlations allows us to recast Eq. (3) as:

$$\begin{aligned}\overline{C'\Gamma C'} &= \langle C'\Gamma C' \rangle \\ &= \int_H \int_H \int_{\Psi(t)} f_2(h,h'|t)C'(h)\Gamma(-t)C'(h')dhdh'dt\end{aligned} \tag{4}$$

where $\langle\rangle$ represents an ensemble average (these replace the volume averages in Eq. (3) by virtue of the ergodic hypothesis), $f_2(h,h'|t)$ denotes the 2-point spatial correlations (i.e., the conditional probability of finding local states h and h' at the tail and head, respectively, of a randomly placed vector t into the microstructure), and $\Psi(t)$ denotes the vector space comprising all of the vectors that can be physically placed in the microstructure volume Ω. In the context of this paper, the local state identifies the distinct thermodynamic phases present in the microstructure of the composite material system (for the inclusion/steel system studied here the local states of interest would simply be the inclusion phase and the steel matrix phase).

The most important and remarkable feature of Eq. (4) is that the contribution of the microstructure to the effective property comes exclusively from the 2-point spatial correlations, for the selected term in the series expansion. Indeed, it can be shown that every term in the series shown in Eq. (2) can be expressed similarly where the role of the material structure enters the homogenization relationship exclusively through a specific set of n-point spatial correlations. For example, the third term in the series of Eq. (2) only needs 3-point spatial correlations as input from the material structure. Note that the very basic set of n-point statistics would correspond to $n = 1$ (referred to as 1-point statistics), and essentially reflect the volume fractions of the different phases in the microstructure. The 1-point statistics enter Eq. (2) through the first term.

Although the strategy described above effectively decomposes the microstructure information from the underlying physics (captured in terms such as the $C'(h)\Gamma(-t)C'(h')$ in Eq. (4)), the practical utility of Eq. (4) is somewhat limited by the complex integrals present [1,4,57]. This is where a digital representation of the microstructure [1,20,31] affords us the next major advance in our endeavor. In this approach, we start with a discretized description of the microstructure as m_s^h, which reflects the volume fraction of local state h residing in the spatial bin s. This is accomplished by binning (typically a uniform tessellation) the spatial domain as well as the local state space. In this notation, $s = 1,2\ldots S$ and $h = 1,2\ldots H$ enumerate the individual bins in the spatial domain and the local state space, respectively. Since the discretized microstructure signal, m_s^h represents the total volume fraction of all local states from bin h residing in the spatial bin s, it exhibits the following properties:

$$\sum_{h=1}^{H} m_s^h = 1, \quad 0 \le m_s^h \le 1 \tag{5}$$

In the same discretized representation, the 2-point correlations are expressed as:

$$f_r^{hh'} = \frac{1}{S}\sum_{s=1}^{S} m_s^h m_{s+r}^{h'} \tag{6}$$

where r enumerates the discretization of the space of vectors using the same uniform binning scheme that was used for binning the spatial domain. For a binary microstructure ($H = 2$), the auto-correlation f_r^{11} captures all of the independent 2-point statistics [1,59].

Employing digital representations allows us to recast the term in Eq. (4) simply as:

$$\langle C'\Gamma C' \rangle = \sum_{r,h,h'} A_r^{hh'} f_r^{hh'} \tag{7}$$

where $f_r^{hh'}$ captures the microstructure information (in the form of 2-point spatial correlations) and the microstructure-independent $A_r^{hh'}$ captures the underlying physics governing the multiscale phenomenon being investigated.

There are a few limitations impeding the practical implementation of the statistical homogenization theories in various multiscale material modeling efforts. It is important to recognize that the number of terms involved in the summation shown in Eq. (7) is extremely large. Although we might expect a number of redundancies in the expression of the n-point spatial correlations, these have not yet been precisely identified [20,59].

In recent work, it was demonstrated that techniques such as Principal Component Analyses (PCA), can be used to obtain objective, orthogonal, prioritized, low-dimensional representations of the 2-point statistics in a selected ensemble of microstructures based on the largest variances present in the dataset [28,30]. The PCA representations of the n-point statistics have allowed an automated classification of an ensemble of microstructures. Since PCA is essentially a linear transformation, one can see that its usage will lead to recasting Eq. (7) as:

$$\langle C'\Gamma C' \rangle \approx \tilde{A}_o + \sum_i \tilde{A}_i \alpha_i \tag{8}$$

where α_i denotes the weights (also called scores) of the principal components of the spatial correlations (details of their computation presented later) and $\tilde{A}_i$ denotes their corresponding influences on the effective property of interest. Physically, α_i captures the most salient information on the microstructure, selected based on the most dominant variances in the spatial correlations in a given ensemble of microstructures. Because Eq. (8) represents essentially a re-mapping of the linkages in Eq. (7), the influence coefficients $\tilde{A}_i$ would still be independent of the material microstructure. Also implicit in Eq. (8) is the truncation

of the series to include only the dominant terms (those making significant contributions to a selected effective property), selected in an objective data-driven approach.

Although the derivations above were discussed by considering only one of the terms in the series of Eq. (2), we hope the reader can see that the above treatment can be applied in the exact same manner to all the terms in the series. However, in the higher-order terms of the series, we need to include higher-order spatial statistics. Since α_i is being used here to denote the weights of the principal components arising from a consideration of all of the n-point spatial correlations included in the analyses, the low-dimensional data-driven form of Eq. (2) looks identical to that shown in Eq. (8):

$$C^* \approx \tilde{A}_o + \sum_i \tilde{A}_i \alpha_i \tag{9}$$

where C^* denotes a specific tensorial component of the effective property of interest.

The goals of the data science approach are to identify the important terms that make dominant contributions to Eq. (9) and estimate the corresponding values of the influence coefficients, $\tilde{A}_i$. Our goal is to keep the structure of Eq. (9) relatively simple with a small number of terms so that we can produce computationally low-cost structure–property linkages. It is also important to note that the relatively simple algebraic structure of Eq. (9) allows us to potentially invert the relationships, where we seek to identify specific microstructures that correspond to a targeted value of the effective property.

One of the limitations of Eq. (9) is that it is derived from the weak contrast expansion of the homogenization theory expressed in Eq. (2). In an effort to alleviate this limitation, we have altered the structure of Eq. (9) as follows:

$$C^* \approx A_o + \sum_i A_i \tilde{\alpha}_i \tag{10}$$

where $\tilde{\alpha}_i$ denotes certain polynomial combinations of α_i. In other words, we have allowed the inclusion of different polynomial groupings of α_i in the surrogate model to be extracted, if they produced significant contributions to the effective property of interest. This extension of Eq. (9) was partially motivated by the fact that the higher-order spatial correlations can be seen as products of lower-order spatial correlations in the limiting case of perfectly disordered materials. Since our interest here extends far beyond perfectly disordered materials, we are hypothesizing that simple polynomials of the weights of the principal components would capture these complex interdependencies to a certain extent.

It should be noted that the introduction of the polynomials in Eq. (10) is a novel contribution of this work. In many ways, this is analogous to Taylor series expansions where one uses similar polynomial expressions. It is also important to note the sequence of operations in the proposed approach. In this approach the PCA is first applied on the spatial correlations without any consideration of the information on the properties. In other words, this is an unsupervised dimensionality reduction based exclusively on structure metrics. Since there was no consideration of properties in the PCA step, there is no implicit assumption of a linear dependence between PC scores and properties. That is why we are seeking nonlinear linkages between PC scores (representing structure) and the properties of interest through Eq. (10). It is also entirely possible to employ other approaches that combine these two steps in a single supervised, nonlinear, dimensionality reduction step. This could be a focus for future efforts in this emerging area.

In the data-driven approach presented in this paper, our strategy would be to objectively evaluate the importance of a large number of structure measures in controlling the value of a selected effective property of interest and identify the specific terms that make the strongest contributions. This is accomplished by performing a large number of regressions to calibration dataset produced using a numerical simulation tool (e.g., finite element models). Data science approaches similar to these have also been successfully employed by our research group in different, but related, applications [60–65].

Overall, the novel data science approach described in this paper involves the three main steps identified in Fig. 1: (1) generating a calibration dataset comprising an ensemble of representative microstructures of interest and simulating their mechanical responses using suitable physics-based numerical models, (2) extracting objective, reduced-order, quantitative measures of the microstructures, and (3) establishing validated structure–property linkages. These steps are described in detail below.

3.1. Generation of the calibration dataset

The first step involves the creation of an ensemble of microstructures that are representative of the specific application, and evaluating their mechanical responses using suitable tools (e.g., finite element models). The microstructures can either be extracted from measurements or generated synthetically. Either way, the microstructure needs to be discretized and digitized suitably, resulting in m_s^h described earlier (cf. [1,31]). The properties of interest corresponding to each microstructure can be obtained either from direct measurements or from mathematical/numerical models.

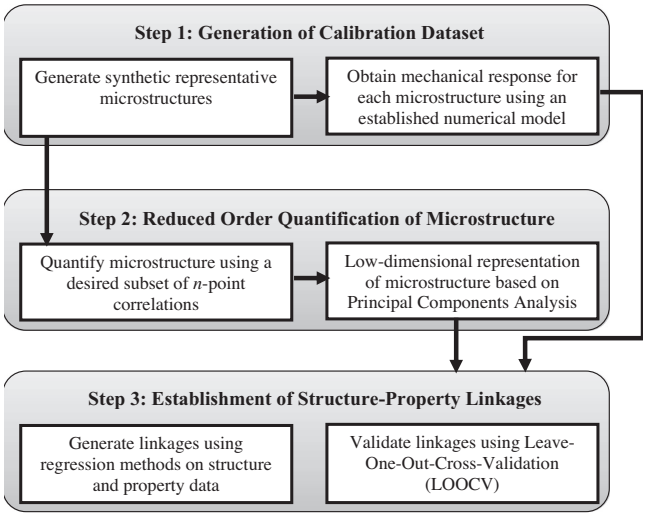

Fig. 1. Main components of the data-science approach utilized in this work for establishing structure–property linkages.

3.2. Establish objective reduced-order quantification of microstructure

Let f_r denote the selected subset of n-point statistics of interest for establishing structure–property linkages in a specific application. Let R denote the dimensionality of f_r, i.e. $r = 1, 2, \ldots, R$. Let $i = 1, 2, \ldots, I$ enumerate the elements of the ensemble of microstructures used in establishing the structure–property linkages of interest. It is generally expected that $I \leq R$. In such situations, PCA identifies a maximum of $(I-1)$ orthogonal directions in the R-dimensional space that are arranged by decreasing levels of variance. Mathematically, the PCA representation of any member of the selected ensemble (of microstructures), labeled by superscript (k), can be expressed as:

$$f_r^{(k)} = \sum_{i=1}^{\min((I-1),R)} \alpha_i^{(k)} \varphi_{ir} + \tilde{f}_r \tag{11}$$

where $\tilde{f}_r$ is simply the averaged value of the selected n-point statistics for the entire ensemble, and $\alpha_i^{(k)}$ (referred to as PC weights) provides an objective representation of the $(k)^{th}$ microstructure in the new orthogonal reference frame identified by φ_{ir} (from PCA).

Another important output from the PCA is the significance of each principal component, b_i, obtained in the eigenvalue decomposition performed as a part of the PCA. The values of b_i provide important measures of the inherent variance among the members of the ensemble of microstructures [30]. More importantly, by retaining only the components associated with the highest (or the most significant) eigenvalues, it is often possible to obtain an objective reduced-order representation of the microstructure with only a handful of parameters [28,30]. Mathematically, this reduced-order representation can be expressed as:

$$f_r^{(k)} \approx \sum_{i=1}^{R^*} \alpha_i^{(k)} \varphi_{ir} + \tilde{f}_r \tag{12}$$

where $R^* \ll \min((I-1), R)$. Selection of R^* will depend on the specific properties that need to be correlated to the microstructure measures.

3.3. Establish and validate structure–property linkages

The protocols described in the previous two sub-sections produce one data point for each microstructure, which can be expressed as $(P_1^{(k)}, P_2^{(k)}, \ldots, P_M^{(k)}, \alpha_1^{(k)}, \alpha_2^{(k)}, \ldots, \alpha_{R^*}^{(k)})$, where (k) indexes the specific microstructure, $P_i^{(k)}$ denotes a specific macroscale property of interest (we assume that there are M macroscale properties of interest), and $\alpha_i^{(k)}$ denotes the reduced-order representation of microstructure (elaborated in the previous subsection). Consider a dataset with K (i.e., $k = 1, 2, \ldots, K$) such data points. We now explore robust methods to extract high fidelity microstructure–property linkages from such a dataset. In this work, we have utilized simple polynomial functions and ordinary least squares linear regression techniques [66]. Other functional descriptions and regression techniques may also be employed as needed.

For simplicity of notation, in the equations below, we drop the subscript on the property of interest and refer to it generically as $P^{(k)}$. Let a normalized error, $E^{(k)}$, associated with each data point in the linear regression be defined as:

$$E^{(k)} = \frac{|P^{(k)} - f^p(\alpha_1^{(k)}, \alpha_2^{(k)}, \ldots, \alpha_{R^*}^{(k)})|}{\frac{1}{K}\sum_{k=1}^{K} P^{(k)}} \tag{13}$$

where $f^p(\alpha_1^{(k)}, \alpha_2^{(k)}, \ldots, \alpha_{R^*}^{(k)})$ denotes a p^{th}-order polynomial function, and the normalization factor is selected as the mean value of the property. The polynomial coefficients are then established using standard protocols of minimizing the sum of the squares of the residuals in the entire dataset (including all K data points). Note that the extracted polynomial linkage depends critically on the selection of both p and R^* as well as the error measure (Eq. (13)).

Critical selection of parameters p and R^* is central to the extraction of high-fidelity microstructure–property linkages. Although higher values of p and R^* will always produce a lower value of the error, they do not necessarily increase the fidelity of the extracted linkages. This is because the higher values of p and R^* may lead to overfitting of the linkages, and can produce erroneous estimates in any subsequent application of the linkages to new microstructures (those not included in the regression analyses).

The following specific measures were employed in this study to critically evaluate the robustness of the polynomial-fits:

(i) Mean absolute error of fit defined as:

$$\overline{E} = \frac{1}{K}\sum_{k=1}^{K} E^{(k)} \tag{14}$$

(ii) Standard deviation, σ of error of fit with respect to mean absolute error, defined as:

$$\sigma = \sqrt{\frac{1}{K}\sum_{k=1}^{K} (E^{(k)} - \overline{E})^2} \tag{15}$$

(iii) Mean absolute error $\overline{E}_{\mathrm{CV}}$, and standard deviation σ_{CV}, of Leave-One-Out Cross Validation (LOOCV) error. This model selection method involves the training of a polynomial fit K times, while leaving one data point out of the test set each time. Cross validation error for a single instance $E_{cv}^{(k)}$ is thus defined as:

$$E_{cv}^{(k)} = \frac{|P^{(k)} - f_{[k]}^p(\alpha_1^{(k)}, \alpha_2^{(k)}, \ldots, \alpha_{R^*}^{(k)})|}{\frac{1}{K}\sum_{k=1}^{K} P^{(k)}} \tag{16}$$

where $f_{[k]}^p$ is the polynomial fit obtained by ignoring the k^{th} data point. Given a large K, for an over-fitted polynomial, the exclusion of a single data point will cause significant change in the coefficients, whereas for a good fit this change will be negligible. The measures $\overline{E}_{\mathrm{CV}}$ and σ_{CV} are defined as the mean and standard deviation of the set defined as $\{E_{cv}^{(k)} \forall k \in 1..K\}$

The first two measures of error of fit defined above will show improvement of fit with higher values of p and R^*, whereas the last two measures of error of CV are expected to show a decline in robustness of fit with higher values of p and R^* (indicating over-fit of data). Therefore, a compromise can be made in choosing the best fit based on the

values of the above four measures. If multiple viable fits exist, one might select the least complex one (the one with the lowest number of terms).

It should be noted that we have deviated here from the traditional practice of splitting the data into a calibration (or training) component and a validation component. The conventional approach of hard-splitting of data is usually a good practice in situations where the number of data points far exceeds the number of parameters involved in the training. In the present application, because of the high cost of generating data (each numerical simulation for a given microstructure using a physics-based model generally incurs a significant cost), the above described approach of cross validation is preferred in order to maximize the utilization of the available information in establishing the surrogate model. Given the unimaginably large space of all theoretically possible microstructures and the relatively small number of available datasets, effective utilization of all available datasets takes precedence in the data science approach described in this paper.

4. Case study: non-metallic inclusions/steel composite system

In this section, we critically evaluate the applicability of the data-science approach described above to the non-metallic inclusions/steel composite system. The importance and motivation for this application have been discussed earlier. The details of the application are described below following the three main steps outlined in the previous section.

4.1. Generation of the calibration dataset

4.1.1. Synthetic microstructures

Two-dimensional (2-D) synthetic microstructures were generated containing multiple hard or soft inclusions with different sizes, shapes and spatial configurations in a steel matrix. For this purpose, a library of 500 particles (inclusions) was first generated such that this set exhibited a broad range of particle shapes and sizes seen in the published literature [41,42,51,67]. Circular, square, triangular, rhombic and platelet (horizontal and vertical orientation) shapes were included in this particle library. Radii of the circular inclusions were selected to be in the size range of 5–30 μm. For inclusions of other shapes, sizes were selected such that they had comparable areas to those of circular inclusions described above. A total of 900 microstructures were generated by randomly selecting the desired number of particles from the library and placing them in the steel matrix based on four different placement criteria – random arrangement, banded arrangements (i.e., horizontal or vertical bands) and clusters. Overlaps in the spatial placement of the particles were allowed, as this essentially led to particles of complex shapes (see Fig. 2). Additional details for each microstructure class are presented below.

(1) *Random arrangement:* A total of 400 synthetic microstructures were generated using the following sequence of steps: (i) randomly choose the number of particles for each microstructure in the range 1–5, (ii) randomly select the desired number of particles from the library, and (iii) randomly place the particles in the microstructure. Fig. 2(a) shows an example of this class of microstructures.

(2) *Banded arrangements:* A total of 250 microstructures were generated by randomly selecting 20 particles in the size range of 5–20 μm from the library, and placing them randomly in one of the three equally spaced horizontal or vertical bands. The number of microstructures were equally split (125 each) into those displaying horizontal and vertical bands. Fig. 2(b) and (c) shows examples of banded microstructures.

(3) *Clusters:* 250 microstructures were generated with inclusions in clusters. 20 particles in size range of 5–20 μm were randomly selected from the library, and placed randomly in one of the four defined inclusion clusters. Fig. 2(d) shows an example of a clustered microstructure.

The selection of the number of microstructures in each class described above was somewhat arbitrary, and was guided by the LOOCV method described earlier (to be performed in step 3 of the proposed approach). In other words, more microstructures of a class were added if the LOOCV analyses (performed in step 3) indicated sensitivity of the extracted linkages to the accumulated datasets.

It is computationally advantageous to discretize the 2-D microstructure into uniform (square) bins and restrict our attention to the class of eigen microstructures (i.e., each pixel was completely filled by either the inclusion or the steel phase). All the microstructures were assumed to occupy spatial dimensions of 250 μm × 250 μm, which was tessellated into 80 × 80 bins. The area fraction of the inclusions in the microstructure ensemble was varied between 0% and 20% to reflect the local volume fraction of inclusions in the steel matrix encountered in practice.

4.1.2. Micro-mechanical model

The next step is to evaluate the properties of interest associated with each of the 900 microstructures generated in the previous step. In the present study, a FEM (finite element model) based micromechanical model was developed to simulate plane strain compression of the inclusion-steel composite system. Furthermore, since the main purpose of this case study is to establish the feasibility of the data science approach, it was decided to evaluate the mechanical properties of interest using a 2-D micro-mechanical model (the 3-D models would have incurred a substantially higher computational cost). The elements in the finite element mesh were made to correspond exactly to the spatial bins in the digitized microstructure. It is noted that the square-shaped elements introduce jagged interfaces (between the inclusions and the matrix), which are likely to result in significant errors in the local stress (or strain) fields close to the interfaces. However, our interest in the present case study is on specific macroscale performance parameters of interest (described later). A sensitivity study was conducted to ensure that the predictions of the macroscale performance parameters selected for this study were not noticeably influenced by any errors introduced by the use of the simple square-shaped elements. This was accomplished by creating alternate meshes with elements suitably designed to align with the interfaces (non-uniform meshes) for a limited number of selected exemplar microstructures, and comparing their predictions with the corresponding predictions from uniform meshes. The use of the simple square-shaped elements allowed automated generation of the finite element meshes for the 900 microstructures used

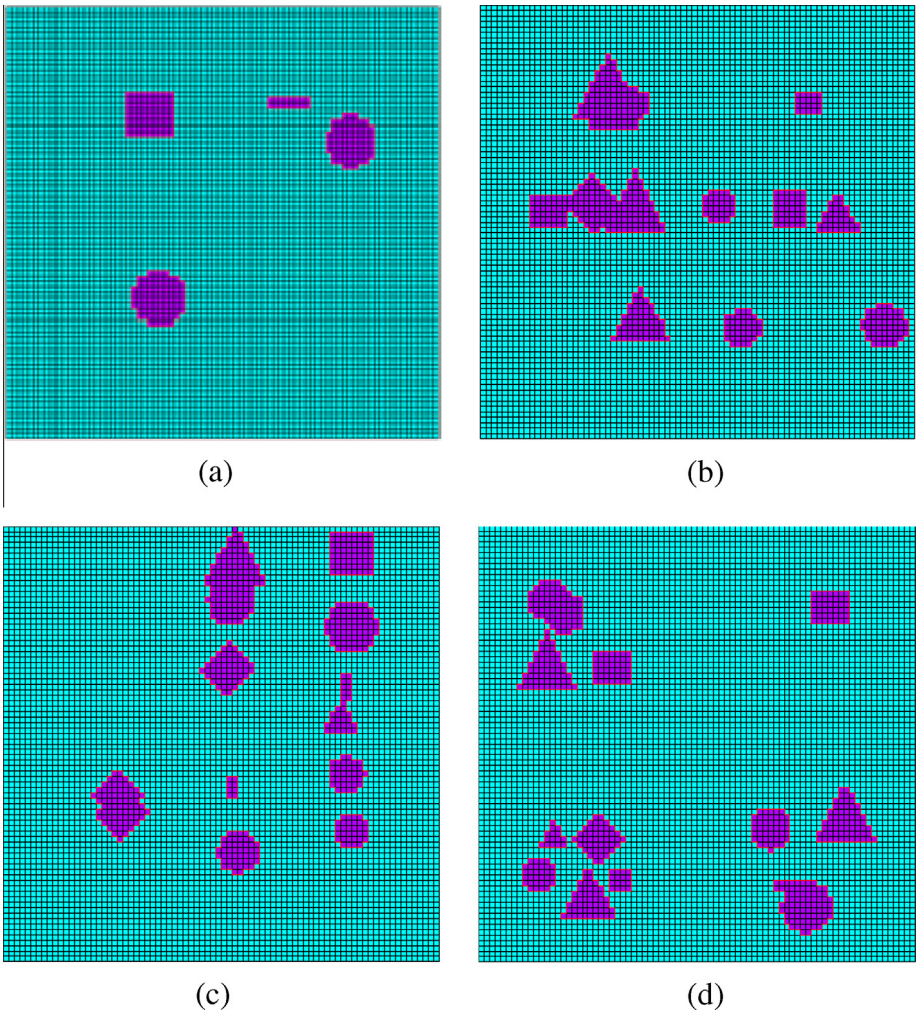

Fig. 2. Examples of the microstructure ensemble generated for this study. Multiple inclusions with different sizes and shapes were randomly selected and distributed in the steel matrix using different placement criteria described in the text.

in this study, and tremendously speeded up the acquisition of the dataset.

The FE model was created using the commercial software ANSYS using 2-D, 4-noded, structural solid elements. A uniform mesh of 6400 elements was generated and periodic boundary conditions were applied to simulate a plane strain compression of 15% reduction. Both the inclusion and steel were treated as isotropic, with the inclusion being elastic-perfectly plastic and the matrix being elastic–plastic (with hardening). The flow stresses for inclusion and steel matrix were prescribed as [68]:

$$\sigma_{\text{inclusion}} = \frac{3}{10}\dot{\varepsilon}T\,\exp\left(\frac{3.39*10^4}{T} - 22.06\right) \tag{17}$$

$$\sigma_{\text{steel}} = 3.37*10^6\,\exp\left(\frac{5000}{T}\right)\varepsilon^{0.23}(0.276\dot{\varepsilon})^{\frac{0.133(T-273)}{1000}} \tag{18}$$

Our attention in the present case study was restricted to temperatures of 900 and 1000 °C and a strain rate of 5 s^{-1} (these reflect the most commonly encountered conditions in hot rolling of steels). Incorporation of a suitable constitutive description for the interfaces is a major challenge [18,49–51]. In this work, the inclusion/steel matrix interface is modeled using 2-D, 2-noded, surface to surface contact elements on the inclusion side of the interface and 2-D target elements on the steel matrix side. It was assumed that the interface is initially bonded, and that sliding occurs with friction after debonding. A bilinear cohesive zone model (CZM) with mode I debonding was used [69]. CZM parameters used in this study included maximum normal contact stress ($\sigma_{max} = 0.01$ MPa), contact gap at the completion of debonding ($\delta_n^c = 0.1$ μm), and artificial damping coefficient ($\eta = 1e{-}5$) [49,70]. Coefficient of friction was taken as 0.4 [50].

A total of three macroscale performance parameters of interest were selected for this case study. In other words for each microstructure, these properties would be extracted from the micro-mechanical FE models and linked with the microstructure measures in the later steps of the protocols described in this paper. The macroscale parameters selected are (i) the effective (composite) yield strength (σ_0), (ii) effective strain hardening exponent (N), and (iii) localization propensity (LP). σ_0 and N are examples of bulk properties (i.e., volume averaged values), while LP quantifies the degree of the heterogeneity of the local response. These parameters are extracted as follows:

(i) σ_0: Effective yield strength of the composite material system is computed from averaged (overall) true stress–strain response. σ_0 is defined as the macroscale stress at 0.2% offset macroscale plastic strain.

(ii) *N:* Effective strain hardening exponent is computed by fitting the averaged true stress–strain response to a power hardening law [70].

$$\sigma_y = \sigma_0\left(1 + \left(\frac{3G}{\sigma_0}\right)\bar{\varepsilon}_p\right)^N \tag{19}$$

In this study, N was computed only for the case of soft inclusions. For the case of hard inclusions, the range of parameter N was negligibly small when computed for all microstructures in the selected ensemble.

(iii) *LP:* Localization propensity was defined to capture strain localization in the system during plane strain compression (to be used as an indicator of potential damage accumulation). It was defined as the area fraction of the matrix elements experiencing an equivalent strain greater than a prescribed cut-off strain (ε_c). In this work, the cut-off strain was chosen as 1.25 times the macroscale equivalent strain ($\bar{\varepsilon}$) applied to the composite material system. It was observed that the predicted values of LP produced a significant variation only for the hard inclusions. This is consistent with prior reports in the literature where hard inclusions were observed to lead to interface delamination, voiding, and eventual crack formation [41–45,48].

Various other parameters used in the models are summarized in Table 1. Separate simulations were performed for (1) hard and (2) soft inclusions, i.e., 900 simulations were performed for each case. Fig. 3(a) and (b) shows example results from simulations for hard and soft inclusions, respectively. The corresponding initial microstructure for these results was shown in Fig. 2(a). Fig. 3(a) shows a contour plot of the equivalent plastic strain for multiple hard inclusions present in the steel matrix. Strain localization and void formation at inclusion/matrix interface can be clearly observed in the simulation results (consistent with experiments reported in the literature [41–45,48]). Fig. 3(b) shows contour plot of equivalent plastic strain for multiple soft inclusions present in the steel matrix. No void formation occurs in the case of soft inclusions; instead inclusions just get elongated.

4.2. Objective reduced-order quantification of microstructure

The next step after generating the calibration dataset is to establish an objective reduced-order quantification of each synthetic microstructure used in the study. In the present study, $\{f_r | r = 1, 2, \ldots, R\}$ denotes the complete set of 2-point statistics for one microstructure. Also, in the present example, $R = 6400$ (in the FFT methods described here, the number of discrete vectors is equal to number of spatial bins, i.e. $80 \times 80 = 6400$). Consequently, the range for the index r is the same as the range for index s used to enumerate the spatial bins. Autocorrelations of the matrix phase were used to capture all of the independent 2-point statistics for each microstructure. 2-point autocorrelations were generated for all 900 microstructures. Following Eq. (11), PCA was performed to obtain reduced-order representations for this set of autocorrelations. Since the number of microstructures was 900, the data matrix of 2-point statistics for PCA was of dimensions 6400×900.

Our first objective with the PCA representations is to examine their efficacy in providing low-dimensional representations of the microstructures. In other words, if two microstructures look similar to each other, their PCA representations (i.e., values of α_i) should be close to each other. Also, if two microstructures look very different, their PCA representations should be far apart. Various two-dimensional projections of the PCA representations (for all of the 900 microstructures included in this study) are shown in Fig. 4(a), where each point in the plot represents one microstructure. It is extremely important to note that all of the plots in Fig. 4(a) are low-dimensional projections of the same PCA space. The microstructure representations in this plot are color-coded to distinguish the four different classes included in this study (i.e., random, horizontally banded, vertically banded, and clustered). Careful examination of the plots shown in Fig. 4(a) does indeed demonstrate that the PCA representations of microstructures within a given class are indeed generally closer to each other compared to PCA representations of microstructures across the different classes. Note that the PCA performed here was completely unsupervised (i.e., the microstructures were not identified by their class in any way in the PCA). Therefore, the automatic classification of the microstructures seen in Fig. 4(a) is actually an output from the PCA. As a further investigation of the efficacy of the PCA, we track the positions of five exemplar microstructures (labeled #1 through #5) shown in Fig. 4(b) in the plots of Fig. 4(a). Microstructures #1 and #2 in Fig. 4(b) are somewhat similar to each other, while being distinctly different than either #4 or #5. Likewise, #4 and #5 are closer to each other, compared to the rest. Also, microstructure #3 is distinctly different from the rest. All of these features are reflected fairly accurately in the distances between their corresponding representations in Fig. 4(a).

It is also informative to examine the information embedded in each of the principal components. Plots of f_r and φ_{ir} (for different values of i; see Eq. (12)) are presented on 80×80 grid of spatial bins in Fig. 5. As noted earlier, r indexes the discretization of the vector space used in defining the 2-point spatial correlations. Plotting the correlations as shown in Fig. 5 allows easy interpretation of the spatial correlations. The center point in these plots corresponds to the zero vector (the corresponding autocorrelation captures the volume fraction). The other points in the plot provide the 2-point statistic of interest for a vector corresponding to the difference between the selected point (where the statistic is shown in the plot) and the center of the plot [26, 27, 29, 30]. In these autocorrelation plots, the intensity and position of each peak with the center provide information on the salient features in the spatial distribution of the matrix phase. The plot of f_r simply reflects the averaged auto-correlations for the entire ensemble of 900 microstructures studied here. As expected the averaged auto-correlation reflects a mostly random microstructure (produces a peak in the center that asymptotes quickly to a uniform value away from the center) with a few small peaks at the midpoints of the edges and the corners (these are presumably the contributions from the banded microstructures). On the other hand, plots of φ_{ir} reflect a prioritized set of orthogonal deviations from the averaged autocorrelation. In other words, φ_{1r} reflects the most dominant deviation, φ_{2r} is the next most dominant

Table 1. Model parameters used for the case study presented in this work.

Definitions	Case 1 (hard inclusions)	Case 2 (soft inclusions)
Temperature (T)	900 °C	1000 °C
Strain rate ($\dot{\varepsilon}$)	5 s^{-1}	5 s^{-1}
Inclusion: Young's modulus (E)	136.93 GPa	10.27 GPa
Inclusion: Poisson's ratio (υ)	0.3	0.3
Inclusion: yield strength (σ_0)	273.86 MPa	20.54 MPa
Steel: Young's modulus (E)	29.77 GPa	21.39 GPa
Steel: Poisson's ratio (υ)	0.3	0.3
Steel: yield strength (σ_0)	56.99 MPa	44.73 MPa
Steel: hardening exponent (N)	0.23	0.23

deviation, and so on. The plot of φ_{1r} shows that it captures a certain scaled deviation in the intensities of the center peak (corresponds to volume fraction of the matrix phase in the microstructure) as well as the peaks at the midpoints of edges and the corners. Note also that the higher-order φ_{ir} are in general more heterogeneous (i.e., they are trying to capture more subtle features in the microstructure). As implied in Eq. (12), one can construct the autocorrelation of any specific microstructure element of interest by starting with the averaged autocorrelation and adding weighted contributions from each of the principal components. This is illustrated in Fig. 6, where the deviation of the autocorrelation of an exemplar microstructure from the ensemble average is decomposed into the individual weighted principal components. As shown in Fig. 6, there will be a truncation error in the decomposition when the higher-order principal components are ignored. However, since PCA provides a prioritized list of principal components, one can make the decision on the truncation level in a very objective manner.

4.3. Establish and validate structure–property linkages

In this study, the structure–property linkages will be generated using both the conventional and the data science approaches described earlier to allow for a critical comparison of these approaches. Structure–property linkages using conventional approaches were established using the following sequence of steps: (1) volume fraction of inclusions and the average inclusion size were computed as structure measures for each synthetic microstructure generated in this study. Volume fraction was computed by dividing the total area of inclusions by total area (inclusion + matrix) and average inclusion size was computed as a simple average of the equivalent diameters of all inclusions present in a microstructure. (2) Macro-scale performance parameters were computed using the micromechanical finite element model (same as for the data-science approach). (3) Power-law functional forms (discussed in Section 2) of the type

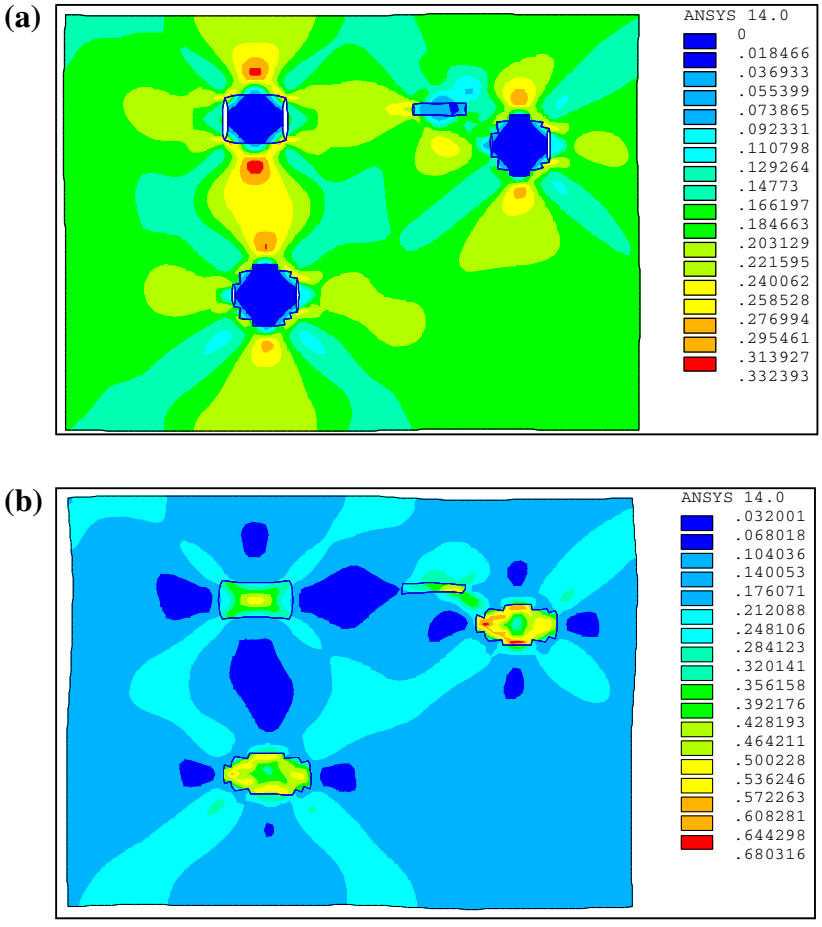

Fig. 3. Typical simulation results showing contour plots of equivalent plastic strain for (a) hard inclusions (case 1), and (b) soft inclusions (case 2) distributed in steel matrix.

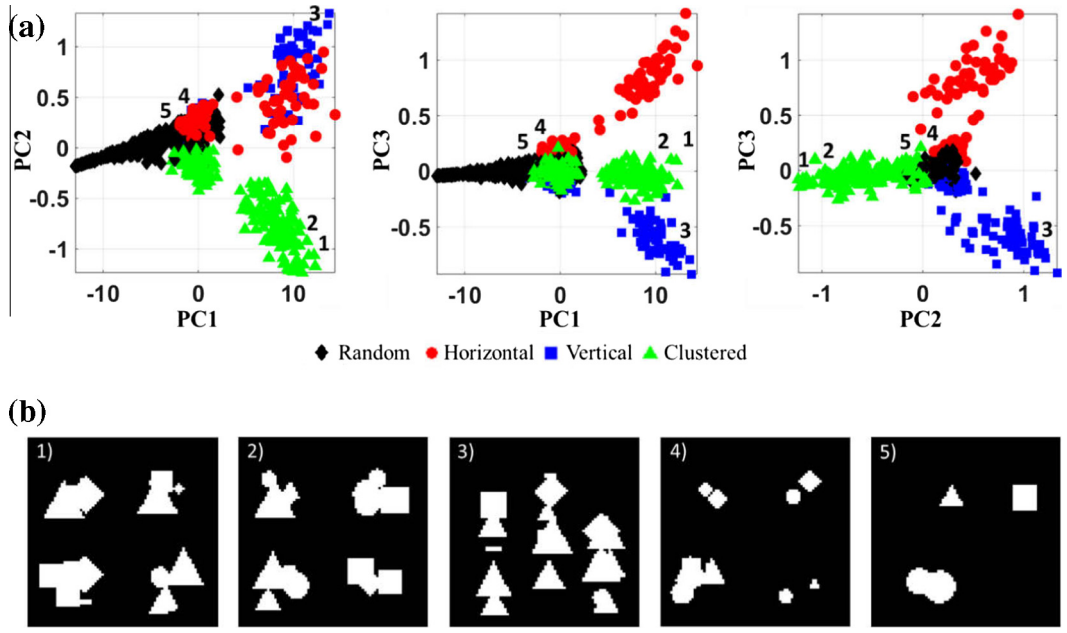

Fig. 4. (a) Scatter plot of the entire data set of 900 micrographs, plotted on the principal components, in different low-dimensional projections. Clusters corresponding to the 4 different methods of microstructure generation are identified by colors and markers. (b) 5 example micrographs whose approximate locations on the first 3 PCA axes are shown in (a). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

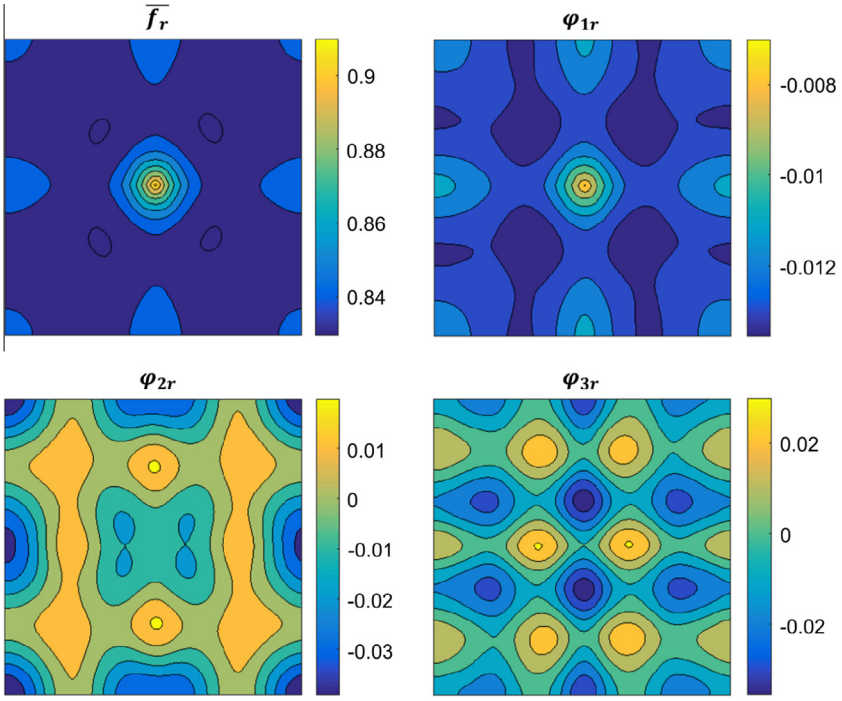

Fig. 5. Plots of $\tilde{f}_r$ and φ_{ir} (for the first 3 values of i; see Eq. (12) and index $r = 1, 2, \ldots, 6400$) obtained from the entire ensemble of 900 microstructures and represented on 80×80 grid of spatial bins.

$y = a^* x^{\lambda}$, where $y =$ property of interest and $x =$ empirical structure measure, were established using standard regression analysis. Following the typical practice in the field, only one structure parameter was explored at a time in the conventional fits.

Structure–property linkages using data science approach were established using regression analysis and validated using leave-one-out cross-validation as described in Section 3.3. Increasing degrees of polynomial ($1 \leqslant p \leqslant 5$) and increasing numbers of PCs ($1 \leqslant R^* \leqslant 5$) were explored in the regression analysis. It should therefore be noted that the data science approach evaluates a very large number of regressions (corresponding to a large number of combinations of p and R^*) before arriving at the best surrogate model. As described earlier, macroscale parameters σ_0, N, and LP were chosen as properties of interest $P_i^{(k)}$. In the

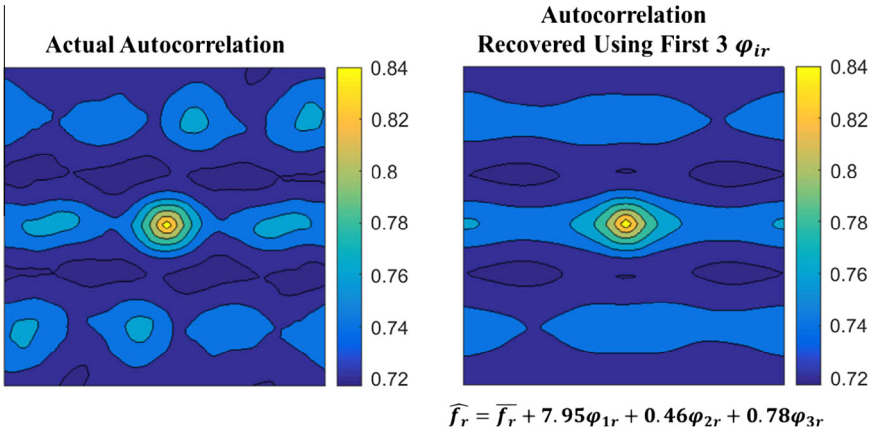

Fig. 6. Comparison of the actual autocorrelation and the PCA approximated autocorrelation for a selected exemplar horizontally banded microstructure.

case of hard inclusions (case 1) parameters σ_0 and LP were computed, and in the case of soft inclusions (case 2) parameters σ_0 and N were computed. Data-driven structure–property linkages for each parameter were then established by objective selection of p and R^* such that the most robust (i.e., accurate without risking over-fitting) linkages were established (after evaluating a large number of regression analyses involving different combinations of p and R^*).

Fig. 7 and Table 2 summarize the results of all the fits produced in this study using both the conventional approaches and the data science approaches described in this paper. It is clear from these results that for all three macroscale parameters studied here, the data science protocols produced very robust and reliable linkages. This is evidenced in the values of the errors and their standard deviations for all 900 microstructures (see Table 2) evaluated in this study. As expected, the linkages are much more accurate for macroscale parameters that reflect bulk averaged values (such as yield strengths) when compared to macroscale parameters that are strongly weighted by local responses (such as LP).

Anyone wanting to use these linkages needs both the values of A_i as well as the corresponding ϕ_{ir}. The size of ϕ_{ir} makes it impossible to present the linkages established in this work in the form of a simple table. The authors would however be willing to share the ϕ_{ir} established in this work with anyone who requests for this information. The values of A_i for each linkage produced in this work are shown in Table 3. The physical significance of the principal components (cf. Fig. 5) was discussed earlier. The linkages presented here allow for quantitative evaluation of the relative influences of any selected principal component or any specific spatial correlation of interest. For example, one might use the linkages shown in Table 3 together with the Eq. (12) to identify the structure features (defined in the space of spatial correlations or in the space of the principal components) that demonstrate the highest potential for improvement in the desired properties (e.g., through evaluation of suitable derivatives).

5. Discussion

Even though Fig. 7 and Table 2 have clearly demonstrated a higher accuracy of the surrogate models obtained from the data science approaches compared to the conventional approaches, this is not the main point of this paper. Indeed, one can argue that the effort involved in establishing the data science linkages presented here is substantially higher compared to the effort expended in the highly simplified conventional approaches presented in this paper. One might even argue that if a stronger effort was spent on the conventional approaches presented here (for example, by combining the different traditional structure measures into new polynomial terms), they might have resulted in equally good (or perhaps even better) surrogate models. However, the main difficulty in extending the conventional approaches to much richer regression analyses comes from the lack of guidance in the objective selection of the structure measures. The data science approach described here, based on the systematic use of n-point statistics for structure measures, presents an approach that can be largely automated using computer codes. It should also be recognized that the overall modular approach presented in Fig. 1 is highly flexible to allow inclusion of any microstructure measures deemed important for the problem at hand, including the simple microstructure measures used here in the conventional approaches. In other words, it is entirely possible to combine the systematic measures of microstructure defined by the n-point statistics with any deemed important by legacy knowledge (i.e., knowledge already accumulated by experts in the specific domain) in the workflow described in Fig. 1. The main point of this paper is to demonstrate that explorations for structure–property linkages can be conducted most efficiently by exploiting modern data science based workflows; the central feature of these new protocols is their automated consideration of a very large number of regression fits leading to the selection of surrogate models that meet the user-prescribed error and cross-validation criteria.

Another distinguishing feature of the approach presented here is that it blends the known physics and data sciences in a very natural manner. Unlike approaches based on neural networks that search through a large space of potential functions for capturing the linkages of interest, the functional form of the linkage in the protocols presented here was taken directly from the most sophisticated homogenization theories available today. Therefore, in the approach presented here, we are effectively combining the best aspects of physics-based approaches and the data science approaches. More specifically, data science was employed to mainly fill the gaps (note that the analytical

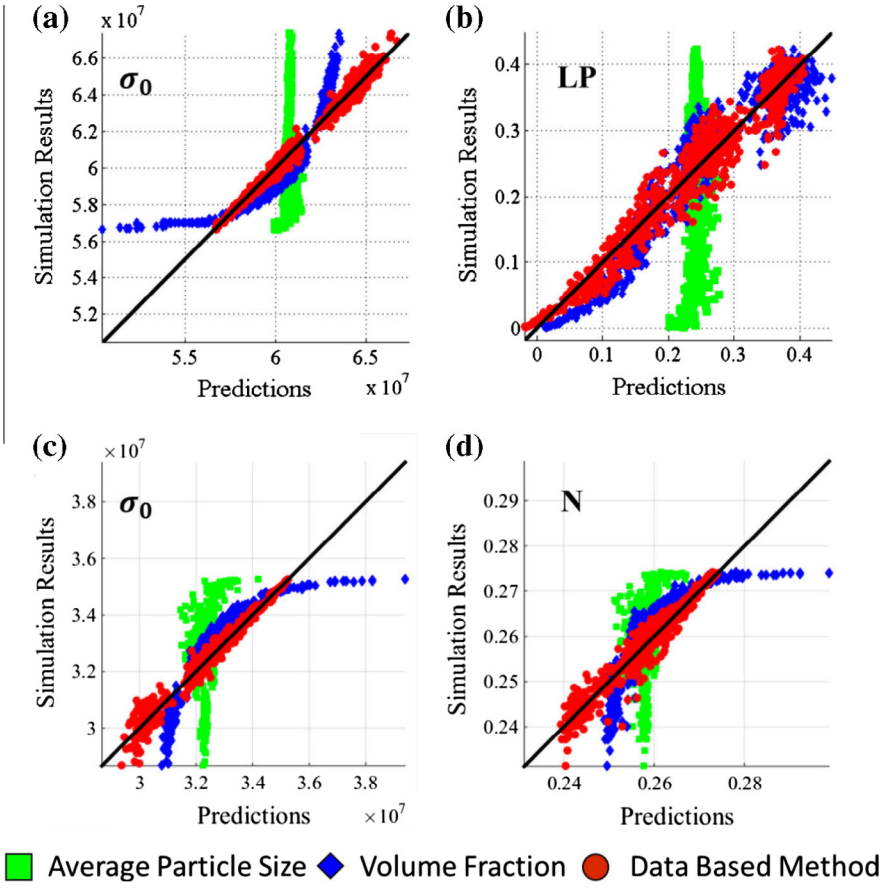

Fig. 7. Comparison of the structure–property linkages using the conventional methods and the data science approaches for the properties/performance characteristics of interest. (a) and (b) Present results for hard inclusions, while (c) and (d) present the results for soft inclusions. The black line represents a perfect match between the estimates from the different surrogate models (extracted using both traditional and data sciences approaches) and the corresponding simulation results.

computation of the influence coefficients, $A_r^{hh'}$ in Eq. (7), is quite challenging).

We now discuss some of the limitations of the data science approach presented here. It is noted that the linkages developed here are data-driven in the sense that they are established using a specific set of 900 microstructures. The central assumption in the data-driven approach presented here is that this set of 900 microstructures is inclusive of all of the different types of microstructures encountered in the actual application. Keeping in mind that the number of theoretically possible distinct microstructures in a composite system is unimaginably large, it is unlikely that the specific set of 900 microstructures selected here would represent every theoretically possible microstructure. In other words, it is entirely possible that one might encounter a new microstructure that is distinctly different from the microstructures included in the study. In such a situation, the framework presented above presents two distinct opportunities: (i) one can quantify rigorously the difference between the new microstructure and the elements of the ensemble used in generating the linkages using the n-point spatial correlations as the measures of the microstructure. It would then be possible to decide objectively if the new microstructure constitutes an interpolation or an extrapolation of the microstructures included in the analyses (cf. see Fig. 4). This is important because an interpolation in the microstructure space would impart much more confidence in the predicted value of the property compared to an extrapolation. (ii) If it is deemed that the new microstructure is distinctly different than the microstructures included in the prior analyses, one can then decide to extend the analyses with minimal additional effort by adding new sets of microstructures. The data science framework and algorithms presented here allow easy addition of new data points with minimal redundant work. This is because both the principal component analyses and the regression methods allow easy updates with the addition of new data points.

There are several areas where the data science approach presented above needs further development and refinement. These include: (i) better approaches for selecting the set of microstructures to be included in the analyses such that they uniformly populate the input domain of interest in establishing the structure–property linkages, and (ii) more rigorous tools for objectively evaluating the relative importance of the exceedingly large number of potential structure measures.

6. Conclusions

The main steps involved in the data-science approach were demonstrated and validated for the case study of non-metallic inclusions/steel composite system. This

Table 2. Summary of the performances of the conventional and data science approaches for establishing structure–property linkages.

		Data based method					Conventional linkages (volume fraction)			Conventional linkages (average particle size)		
		# PCs	Degree	R^2	$\bar{E}$ %	σ %	R^2	$\bar{E}$ %	σ %	R^2	$\bar{E}$ %	σ %
Hard inclusions	σ_0	1	2	0.9847	0.38	0.34	0.7370	1.68	1.25	0.0068	3.36	2.31
	LP	2	3	0.9503	8.18	6.64	0.9004	12.48	8.18	0.0111	39.98	24.74
Soft inclusions	σ_0	1	2	0.9708	0.58	0.55	0.7563	1.82	1.43	0.0370	3.93	2.41
	N	3	2	0.9451	0.60	0.59	0.6794	1.62	1.24	0.0593	2.91	1.94

Table 3. Coefficients for the polynomial fits obtained through the data science approach for each parameter explored in this study.

Polynomial terms	Coefficients of polynomial terms			
	Hard inclusions		Soft inclusions	
	σ_0	LP	σ_0	N
Constant	6.06E+7	0.28	3.22E+7	0.26
PC1	6.89E+7	3.63	−3.72E+7	−0.23
PC2		−0.52		0.03
PC3				−0.10
$PC1^2$	1.74E+8	−26.98	6.37E+7	−0.04
$PC2^2$		−7.27		−0.07
$PC3^2$				0.25
PC1*PC2		17.44		−0.41
PC1*PC3				0.85
PC2*PC3				0.06
$PC1^3$		−145.83		
$PC2^3$		−54.01		
$PC1^2$*PC2		−97.58		
$PC2^2$*PC1		130.85		

approach was evaluated for the first time for nonlinear phenomenon such as plasticity, and the good fits obtained attest to the versatility of this approach. Since, inclusions have different shapes, sizes and distributions in the steel matrix, a very large number of different microstructures could be potentially encountered in this application. The data-science approach provided a practical approach to this problem and extracted robust structure–property linkages of interest. These structure–property linkages are capable of communicating systematically the knowledge gathered at microscale to macroscale simulations, and are invaluable to process design.

Acknowledgements

Akash Gupta, Sharad Goyal and A.K. Singh acknowledge encouragement and support from TCS CTO, Mr. K Ananth Krishnan, and TRDDC Process Engineering Lab Head, Dr. Pradip. Ahmet Cecen and Surya Kalidindi acknowledge support from AFOSR award FA9550-12-1-0458.

Appendix A

List of important symbols and their definitions used in this paper.

Symbol	Definition	Symbol	Definition
C^*	Effective stiffness tensor	b_i	Variance of principal components
C'	Perturbation in local stiffness tensor	P_i	Macroscale property of interest
Ω	Microstructure volume	K	Total number of microstructures
t	Position vector	p	Order of polynomial
h, h'	Local states	f^p	p^{th}-order polynomial function

Ψ	Position vector space	$\overline{E}$	Mean absolute error of fit
s	Index for spatial bin	σ	Standard deviation of error of fit
m_s^h	Discretized microstructure signal	$\overline{E}_{CV}$	Mean absolute error of LOOCV
r	Index for discretized vector space	σ_{CV}	Standard deviation of LOOCV
$f_r^{hh'}$, f_r	2-point spatial correlations	σ_0	Yield strength
$\tilde{f}_r$	Averaged 2-point correlations	N	Strain hardening exponent
α_i	Weights of principal components	LP	Localization propensity
$\tilde{A}_i$, $\tilde{A}_0$	Influence coefficients of α_i	$\overline{\varepsilon}$	Macroscale equivalent strain
$\tilde{\alpha}_i$	Polynomial combinations of α_i	ε_c	Cut-off strain
φ_{ir}	Orthogonal PCA reference frame	k	Index for number of microstructures
A_i, A_0	Influence coefficients of $\tilde{\alpha}_i$	R^*	Number of principal components used for developing linkages
R	Total number of 2-point statistics for a microstructure	$f_r^{(k)}$	2-point statistics for k^{th} microstructure
Γ	Symmetric second derivative of Green's function	$E^{(k)}$	Normalized error for k^{th}microstructure
$A_r^{hh'}$	Microstructure independent coefficients of $f_r^{hh'}$		

References

[1] B.L. Adams, S.R. Kalidindi, D. Fullwood, Microstructure Sensitive Design for Performance Optimization, Butterworth-Heinemann, 2012.

[2] B.L. Adams, T. Olson, Mesostructure–properties linkage in polycrystals, Prog. Mater. Sci. 43 (1998) 1–88.

[3] P.P. Castañeda, J.J. Telega, B. Gambin, Nonlinear Homogenization and Its Applications to Composites, Polycrystals and Smart Materials, Springer, 2004.

[4] D.T. Fullwood, B.L. Adams, S.R. Kalidindi, A strong contrast homogenization formulation for multi-phase anisotropic materials, J. Mech. Phys. Solids 56 (2008) 2287–2297.

[5] D.J. Luscher, D.L. McDowell, C.A. Bronkhorst, A second gradient theoretical framework for hierarchical multiscale modeling of materials, Int. J. Plast. 26 (2010) 1248–1275.

[6] D.L. McDowell, H.J. Choi, J. Panchal, R. Austin, J. Allen, F. Mistree, Plasticity-related microstructure–property relations for materials design, Key Eng. Mater. 340–341 (2007) 21–30.

[7] D.L. McDowell, G.B. Olson, Concurrent design of hierarchical materials and structures, Sci. Model. Simul. 15 (2008) 207–240.

[8] G.W. Milton. The Theory of Composites Cambridge Monographs on Applied and Computational Mathematics (2001).

[9] P. Ponte Castañeda, Second-order homogenization estimates for nonlinear composites incorporating field fluctuations: I—theory, J. Mech. Phys. Solids 50 (2002) 737–757.

[10] P. Ponte Castañeda, Second-order homogenization estimates for nonlinear composites incorporating field fluctuations: II—applications, J. Mech. Phys. Solids 50 (2002) 759–782.

[11] F. Roters, P. Eisenlohr, L. Hantcherli, D.D. Tjahjanto, T.R. Bieler, D. Raabe, Overview of constitutive laws, kinematics, homogenization and multiscale methods in crystal plasticity finite-element modeling: theory, experiments, applications, Acta Mater. 58 (2010) 1152–1211.

[12] D.R.S. Talbot, J.R. Willis, Upper and lower bounds for the overall response of an elastoplastic composite, Mech. Mater. 28 (1998) 1–8.

[13] D.R.S. Talbot, J.R. Willis, Upper and lower bounds for the overall properties of a nonlinear composite dielectric. II. Periodic microgeometry, proceedings: mathematical and physical, Sciences 447 (1994) 385–396.

[14] D.R.S. Talbot, J.R. Willis, V. Nesi, On improving the Hashin-Shtrikman bounds for the effective properties of three-phase composite media, IMA J. Appl. Math. 54 (1995) 97–107.

[15] J.R. Willis, Variational and related methods for the overall properties of composites, Adv. Appl. Mech. 21 (1981) 1–78.

[16] J. Bouvard, D. Ward, D. Hossain, S. Nouranian, E. Marin, M. Horstemeyer, Review of hierarchical multiscale modeling to describe the mechanical behavior of amorphous polymers, J. Eng. Mater. Technol. 131 (2009) 041206.

[17] J. Elliott, Novel approaches to multiscale modelling in materials science, Int. Mater. Rev. 56 (2011) 207–225.

[18] S. Hao, W.K. Liu, B. Moran, F. Vernerey, G.B. Olson, Multi-scale constitutive model and computational framework for the design of ultra-high strength, high toughness steels, Comput. Methods Appl. Mech. Eng. 193 (2004) 1865–1908.

[19] S. Torquato, Random Heterogeneous Materials, Springer-Verlag, New York, 2002.

[20] D.T. Fullwood, S.R. Niezgoda, B.L. Adams, S.R. Kalidindi, Microstructure sensitive design for performance optimization, Prog. Mater. Sci. 55 (2010) 477–562.

[21] T. Baker, K. Gave, J. Charles, Inclusion deformation and toughness anisotropy in hot-rolled steels, Met. Technol. 3 (1976) 183–193.

[22] W.M. Garrison Jr., A.L. Wojcieszynski, A discussion of the effect of inclusion volume fraction on the toughness of steel, Mater. Sci. Eng. A 464 (2007) 321–329.

[23] A. Gupta, S. Goyal, K.A. Padmanabhan, A.K. Singh, Inclusions in steel: micro–macro modelling approach to analyse the effects of inclusions on the properties of steel, Int. J. Adv. Manuf. Technol. (2014), http://dx.doi.org/10.1007/s00170-014-6464-5.

[24] J. Lankford, (E) Effect of oxide inclusions on fatigue failure, Int. Met. Rev. 22 (1977) 221–228.

[25] Y. Murakami, Metal Fatigue: Effects of Small Defects and Nonmetallic Inclusions, Elsevier, 2002.

[26] D.T. Fullwood, S.R. Niezgoda, S.R. Kalidindi, Microstructure reconstructions from 2-point statistics using phase-recovery algorithms, Acta Mater. 56 (2008) 942–948.

[27] S.R. Niezgoda, D.T. Fullwood, S.R. Kalidindi, Delineation of the space of 2-point correlations in a composite material system, Acta Mater. 56 (2008) 5285–5292.
[28] S.R. Kalidindi, S.R. Niezgoda, A.A. Salem, Microstructure informatics using higher-order statistics and efficient data-mining protocols, JOM 63 (2011) 34–41.
[29] S.R. Niezgoda, A.K. Kanjarla, S.R. Kalidindi, Novel microstructure quantification framework for databasing, visualization, and analysis of microstructure data, Integr. Mater. Manuf. Innov. 2 (2013) 1–27.
[30] S.R. Niezgoda, Y.C. Yabansu, S.R. Kalidindi, Understanding and visualizing microstructure and microstructure variance as a stochastic process, Acta Mater. 59 (2011) 6387–6400.
[31] B.L. Adams, X. Gao, S.R. Kalidindi, Finite approximations to the second-order properties closure in single phase polycrystals, Acta Mater. 53 (2005) 3563–3577.
[32] W.F. Brown Jr., Solid mixture permittivities, J. Chem. Phys. 23 (1955) 1514–1517.
[33] S.R. Kalidindi, Microstructure informatics, in: K. Rajan (Ed.), Informatics for Materials Science and Engineering: Data-Driven Discovery for Accelerated Experimentation and Application, Butterworth-Heinemann, 2013, pp. 443–466.
[34] S.R. Kalidindi, Data science and cyberinfrastructure: critical enablers for accelerated development of hierarchical materials, Int. Mater. Rev. 60 (2014) 150–168.
[35] A. Agrawal, P.D. Deshpande, A. Cecen, G.P. Basavarsu, A.N. Choudhary, S.R. Kalidindi, Exploration of data science techniques to predict fatigue strength of steel from composition and processing parameters, Integr. Mater. Manuf. Innov. 3 (2014) 8.
[36] K.D.M. Aris, F. Mustapha, M.S. Salit, D.L.A.A. Majid, Condition structural index using principal component analysis for undamaged, damage and repair conditions of carbon fiber-reinforced plastic laminate, J. Intell. Mater. Syst. Struct. 25 (2014) 575–584.
[37] K. Rajan, C. Suh, P.F. Mendez, Principal component analysis and dimensional analysis as materials informatics tools to reduce dimensionality in materials science and engineering, Stat. Anal. Data Mining 1 (2009) 361–371.
[38] P.M. Shenai, Z. Xu, Y. Zhao, Applications of principal component analysis (PCA) in materials science, in: P. Sanguansat (Ed.), Principal Component Analysis – Engineering Applications, InTech, 2012, pp. 25–40.
[39] S.M. Qidwai, D.M. Turner, S.R. Niezgoda, A.C. Lewis, A.B. Geltmacher, D.J. Rowenhorst, S.R. Kalidindi, Estimating the response of polycrystalline materials using sets of weighted statistical volume elements, Acta Mater. 60 (2012) 5284–5299.
[40] A. Çeçen, T. Fast, E. Kumbur, S. Kalidindi, A data-driven approach to establishing microstructure–property relationships in porous transport layers of polymer electrolyte fuel cells, J. Power Sources 245 (2014) 144–153.
[41] A. Ghosh, Secondary Steelmaking: Principles and Applications, CRC Press, 2000.
[42] H. Jacobi, F. Rakoski. High purity in steels as a criterion for materials development, Le Journal de Physique IV 5 (1995) C7-3-C7-22.
[43] N. Wolanska, A. Lis, J. Lis, Microstructure investigation of low carbon steel after hot deformation, J. Achievements Mater. Manuf. Eng. 20 (2007) 291–294.
[44] K. Yamamoto, H. Yamamura, Y. Suwa, Behavior of non-metallic inclusions in steel during hot deformation and the effects of deformed inclusions on local ductility, ISIJ Int. 51 (2011) 1987–1994.
[45] L. Zhang, B.G. Thomas, Inclusions in Continuous Casting of Steel, XXIV National Steelmaking Symposium, Morelia, Mich, Mexico, 2003, 26–28.
[46] R. Prasannavenkatesan, J. Zhang, D.L. McDowell, G.B. Olson, H.-J. Jou, 3D modeling of subsurface fatigue crack nucleation potency of primary inclusions in heat treated and shot peened martensitic gear steels, Int. J. Fatigue 31 (2009) 1176–1189.
[47] U.T. Riedel, W. Bleck, J.E. Morgan, F.J. Guild, C.A. McMahon, Finite element modelling of the effect of non-metallic inclusions in metal forming processes, Comput. Mater. Sci. 16 (1999) 32–38.
[48] E. Ervasti, U. Ståhlberg, Void initiation close to a macro-inclusion during single pass reductions in the hot rolling of steel slabs: a numerical study, J. Mater. Process. Technol. 170 (2005) 142–150.
[49] C. Luo, Evolution of voids close to an inclusion in hot deformation of metals, Comput. Mater. Sci. 21 (2001) 360–374.
[50] A. Stienon, A. Fazekas, J.-Y. Buffière, A. Vincent, P. Daguier, F. Merchi, A new methodology based on X-ray micro-tomography to estimate stress concentrations around inclusions in high strength steels, Mater. Sci. Eng. A 513 (2009) 376–383.
[51] H.-L. Yu, H.-Y. Bi, X.-H. Liu, L.-Q. Chen, N.-N. Dong, Behavior of inclusions with weak adhesion to strip matrix during rolling using FEM, J. Mater. Process. Technol. 209 (2009) 4274–4280.
[52] E. Kroner, Statistical modelling, in: J. Gittus, J. Zarka (Eds.), Modelling Small Deformations of Polycrystals, Elsevier Science Publishers, London, 1986, pp. 229–291.
[53] M. Beran, J. Molyneux, Statistical properties of the electric field in a medium with small random variations in permittivity, Nuovo Cimento 30 (1963) 1406–1422.
[54] M.J. Beran, T.A. Mason, B.L. Adams, T. Olsen, Bounding elastic constants of an orthotropic polycrystal using measurements of the microstructure, J. Mech. Phys. Solids 44 (1996) 1543–1563.
[55] H. Garmestani, S. Lin, B.L. Adams, Statistical continuum theory for inelastic behavior of a two-phase medium, Int. J. Plast. 14 (1998) 719–731.
[56] H. Garmestani, S. Lin, B.L. Adams, S. Ahzi, Statistical continuum theory for large plastic deformation of polycrystalline materials, J. Mech. Phys. Solids 49 (2001) 589–607.
[57] S.R. Kalidindi, M. Binci, D. Fullwood, B.L. Adams, Elastic properties closures using second-order homogenization theories: case studies in composites of two isotropic constituents, Acta Mater. 54 (2006) 3117–3126.
[58] T.A. Mason, B.L. Adams, Use of microstructural statistics in predicting polycrystalline material properties, Metall. Mater. Trans. A 30 (1999) 969–979.
[59] A.M. Gokhale, A. Tewari, H. Garmestani, Constraints on microstructural two-point correlation functions, Scripta Mater. 53 (2005) 989–993.
[60] T. Fast, S.R. Kalidindi, Formulation and calibration of higher-order elastic localization relationships using the MKS approach, Acta Mater. 59 (2011) 4595–4605.
[61] T. Fast, S.R. Niezgoda, S.R. Kalidindi, A new framework for computationally efficient structure-structure evolution linkages to facilitate high-fidelity scale bridging in multi-scale materials models, Acta Mater. 59 (2011) 699–707.
[62] S.R. Kalidindi, Computationally-efficient fully-coupled multi-scale modeling of materials phenomena using calibrated localization linkages, ISRN Mater. Sci. 2012 (2012), http://dx.doi.org/10.5402/2012/305692. Article ID 305692, 13 pages.
[63] S.R. Kalidindi, S.R. Niezgoda, G. Landi, S. Vachhani, T. Fast, A novel framework for building materials knowledge systems, Comput. Mater. Con. 17 (2010) 103–125.
[64] G. Landi, S.R. Niezgoda, S.R. Kalidindi, Multi-scale modeling of elastic response of three-dimensional voxel-based microstructure datasets using novel DFT-based knowledge systems, Acta Mater. 58 (2010) 2716–2725.
[65] Y.C. Yabansu, D.K. Patel, S.R. Kalidindi, Calibrated localization relationships for elastic response of polycrystalline aggregates, Acta Mater. 81 (2014) 151–160.
[66] C.R. Rao, Linear Statistical Inference and Its Applications, John Wiley & Sons, 2009.
[67] D.X. Yang, J.P. Xie, K.F. Zhang, Z.F. Liu, A.Q. Wang, Numerical simulation of stress field in inclusions of large rudder arm steel castings, China Foundry 6 (2009) 219–225.

[68] G. Bernard, P. Riboud, G. Urbain, Investigation of the plasticity of oxide inclusions, Revue de Metallurgie, Cahiers d'Informations Techniques 78 (1981) 421–433.
[69] G. Alfano, M. Crisfield, Finite element interface models for the delamination analysis of laminated composites: mechanical and computational issues, Int. J. Numer. Meth. Eng. 50 (2001) 1701–1736.
[70] ANSYS Release 14.0, Help System, Mechanical APDL, ANSYS Inc.