Computational Materials Science 156 (2019) 17–25

Contents lists available at ScienceDirect

Computational Materials Science

journal homepage: www.elsevier.com/locate/commatsci

An efficient machine learning approach to establish structure-property linkages

Jaimyun Jung[a], Jae Ik Yoon[a], Hyung Keun Park[a], Jin You Kim[b], Hyoung Seop Kim[a,c,*]

[a] Department of Materials Science and Engineering, Pohang University of Science and Technology (POSTECH), Pohang 37673, Republic of Korea
[b] Pohang Research Lab. Steel Products Research Group 1, POSCO, Pohang 790-785, Republic of Korea
[c] Center for High Entropy Alloys, Pohang University of Science and Technology (POSTECH), Pohang 37673, Republic of Korea

ARTICLE INFO

Keywords:
Microstructure
Machine learning
Gaussian process regression
Optimization

ABSTRACT

Full-field simulations with synthetic microstructure offer unique opportunities in predicting and understanding the linkage between microstructural variables and properties of a material prior to or in conjunction with experimental efforts. Nevertheless, the computational cost restrains the application of full-field simulations in optimizing materials microstructures or in establishing comprehensive structure-property linkages. To address this issue, we propose the use of machine learning technique, namely Gaussian process regression, with a small number of full-field simulation results to construct structure-property linkages that are accurate over a wide range of microstructures. Furthermore, we demonstrate that with the implementation of expected improvement algorithm, microstructures that exhibit most desirable properties can be identified using even smaller number of full-field simulations.

1. Introduction

While traditional trial and error has served well in designing material microstructures with desired or even enhanced properties and performance, computational means to aid microstructure design are highly desirable for further accelerating materials design. Unfortunately, the linkage between microstructure and properties is a vastly complex one. The sheer complexity in quantifying the geometry of the microstructure coupled with the uncertainty brought about by how different phases mechanically interact given specific structures are only two of many frustrations in computationally aided microstructure design.

There are several approaches available for establishing structure-property linkages. Some of the conventional approaches include analytical approach based on statistical continuum theories [1–3], mean-field approach based on Eshelby's inclusion problems [4–6], numerical approach based on finite element method (FEM). The approach based on statistical continuum theories, when successfully established, is computationally very cheap. Nonetheless, the approach is hindered by difficulties regarding the derivation of analytic expressions for Green's function kernels and the convergence of the series expansions employed in the approach [7,8]. Mean-field approach is also known to be computationally cheap. However, the approach is unable to predict stress or strain localization and often cannot take into account complex morphology and spatial distribution of differing phases. The numerical approach involving FEM is often limited by its computational cost. To date, several works that employed a framework that establishes process-structure-property (PSP) linkages using low dimensional representation of microstructures and regression models as surrogate models are available [9–14]. In the framework, microstructures are first quantified using two-point correlation statistics. Then, principal component analysis (PCA) is applied to the statistics to reduce its dimension. Finally, a surrogate model is constructed using the low dimensional representation of microstructures as input variables. Some of these works employed full-field simulations to capture the PSP linkages [9–14]. In particular, Latypov et al. [9] captured structure-property (SP) relations using microstructure based simulation results on synthetic microstructures for multivariate regression.

The advantage of full-field simulations with synthetic microstructures is that these simulations can quantitatively evaluate how microstructural variations affect the mechanical properties of materials [9,15]. This type of modelling approach already demonstrated its effectiveness in describing and predicting mechanical behavior of composites [16], steels [17–20], and other alloys [21–23]. Despite their merits, microstructure based modelling techniques suffer from their high computational cost. In the framework introduced above, the issue

[*] Corresponding author at: Department of Materials Science and Engineering, Pohang University of Science and Technology (POSTECH), Pohang 37673, Republic of Korea.
E-mail address: hskim@postech.ac.kr (H.S. Kim).

https://doi.org/10.1016/j.commatsci.2018.09.034
Received 28 May 2018; Received in revised form 13 September 2018; Accepted 15 September 2018
Available online 21 September 2018

is mainly addressed by constructing an approximate model. Approximate model, or surrogate model, is any model that can mimic the behavior of full-field models at a fraction of their computational cost. Nevertheless, any surrogate model to replace full-field simulations will still require output from the simulations. Consequently, the time it takes to develop an accurate surrogate will be proportional to the number of the simulations necessary. Therefore, robust surrogate modelling with an efficient sampling scheme is necessary to expedite the process of constructing SP linkages over a wide range of microstructures.

New opportunities for surrogate modelling have opened up with the recent surge of interest in machine learning (ML) techniques. Many ML methods such as artificial neural network models [24], Gaussian process (GP) based methods [25], decision-tree and random forest models [26], and kernel-based methods [27] can serve as accurate surrogate models. Among these methods, GP methods offer a principled approach in dealing with model uncertainty. In particular, Gaussian process regression (GPR) can predict unobserved values as well as their uncertainty. This allows one to selectively conduct full-field simulations where high uncertainty in the ML model is expected. Furthermore, GPR is known to be suitable for optimization of expensive black-box simulations via Efficient Global Optimization (EGO) method based on expected improvement (EI) algorithm [28], making GPR a very attractive approach for microstructure optimization using full-field simulations. In this work, we extended the recently developed data-driven frameworks for SP linkages [9–14] by combining synthetic microstructure based simulations and GP based ML to construct SP linkages. Furthermore, we implemented the EI algorithm to find optimal microstructures within a large microstructure database. We first demonstrate effectiveness of the approach in constructing an accurate SP linkages over 1100 synthetic two-phase microstructures using only a fraction of the microstructures for full-field simulations. Afterwards, by implementing the EI algorithm, we show that the approach can search for an optimal microstructure that maximizes specific property within the dataset with even fewer full-field simulations.

2. Methods

2.1. Summary of the method

The overall schematic for the proposed approach is presented in Fig. 1. Firstly, a database consisting of large amount of synthetic microstructures are generated (Section 2.2). The microstructures are then quantified using two-point correlation statistics, which has been successfully used in microstructure quantifications for surrogate models [9–14]. Following the works that establish PSP linkages [9–14], the two-point correlation statistics of synthetic microstructures is projected into a low-dimensional space. In our work, multi-dimensional scaling (MDS) instead of PCA was used as dimensionality reduction technique for better between-data distances preservation (Section 2.3). Microstructure based simulations were conducted with randomly sampled microstructure dataset. The GPR model was trained with the simulated results using low-dimensional microstructural variables as input features. Afterwards, full-field simulations were preferentially conducted with microstructure data with highest predicted variance (Section 2.5), or highest expected improvement for optimization (Section 2.6), until a specified stopping criteria was reached. There are limited work regarding the stopping criteria [29–32]. These works typically utilized 0.001 to 0.002 as stopping criteria. In this study, we adopted a stopping criteria that terminated the GPR training when the maximum predicted variance, or EI, reached below 0.0015. GPflow [33] was used to implement GPR and Matlab 2015a was used to perform MDS and PCA. A Python script was written to implement the overall algorithm shown in Fig. 1.

2.2. Microstructure generation

In order to construct SP linkages over a wide range of microstructural features, a database filled with microstructures reflecting those features is necessary. In this work, a database consisting of two-phase microstructures composed of 27 × 27 × 27 voxels were generated with periodic boundary condition using the open-source software DREAM.3D [34]. The size of the synthetic microstructures were selected based on the works by Latypov et al. [9].

We adopted a similar technique used by Latypov et al. [9] to generate different classes of microstructures by controlling the volume fraction, size distribution, and aspect ratio of second phase inclusions. Because inclusions are not allowed to overlap in the inclusion/matrix microstructures in DREAM.3D, as specified in [9], inclusion/matrix microstructures with inclusion volume fractions in the range of 5% to 35% and 65% to 95% were generated. Equiaxed two phase

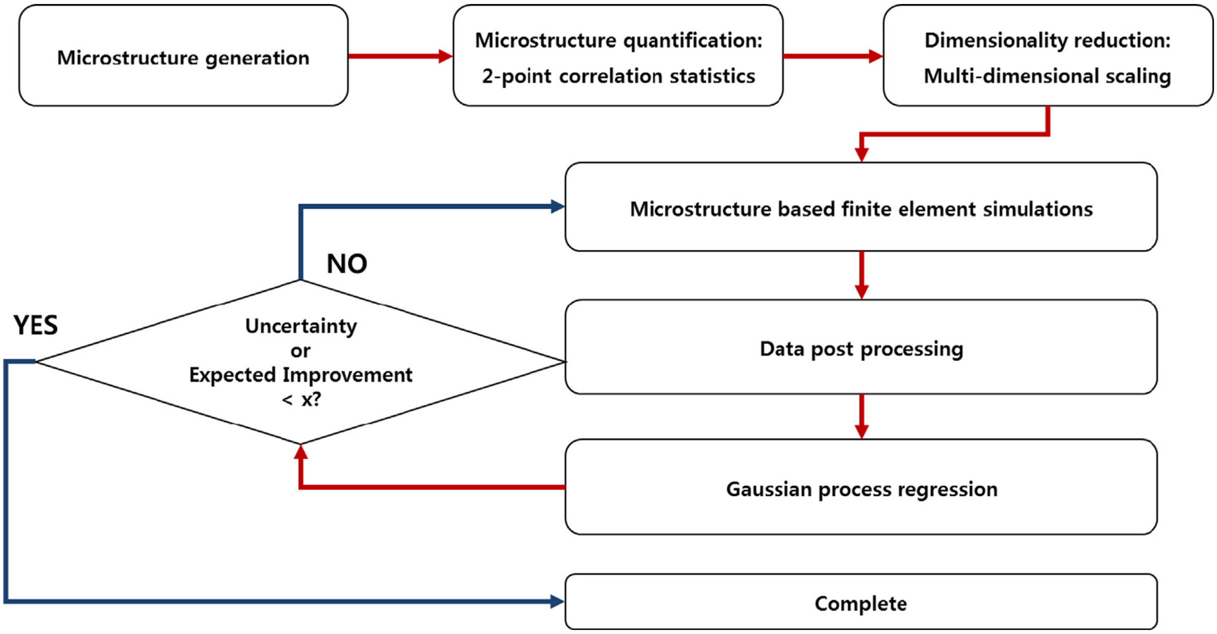
Microstructure generation
Microstructure quantification: 2-point correlation statistics
Dimensionality reduction: Multi-dimensional scaling
Microstructure based finite element simulations
Data post processing
Gaussian process regression
Uncertainty or Expected Improvement < x?
NO
YES
Complete

Fig. 1. Schematic of the proposed framework.

microstructures were generated for microstructures with second phase volume fractions in the range of 35% to 65%. Aspect ratio of 3D inclusion is defined as R_1: R_2: R_3 where R_i is the length of an ellipsoid along i-axis. Microstructures were generated with $R_{i=1,2,3} = 1, 5, 7,$ or 10. where $R_1 \geqslant R_2 \geqslant R_3$. We also allowed elongated inclusions to be either randomly oriented or aligned along a common axis. DREAM.3D defines inclusion size distribution through a log-normal distribution of equivalent sphere diameter, which requires an average value (μ), standard deviation (σ), and minimum (σ_{min}) and maximum (σ_{max}) cut-off values. In this work, μ was set to 0.5, 1.3, and 2.1, and σ was set to 0.1 for $\mu = 0.1$ and 0.5 for the remaining. The cut-off values σ_{min} and σ_{max} were fixed to 3.0 and 1.5, respectively. The resulting microstructure database used in this study contained over 1100 different microstructures.

2.3. Low dimensional representation of microstructure

Microstructure quantification is required to successfully formulate surrogate models. Two-point correlation statistics has been successful in quantifying microstructures as well as formulating surrogate models for SP linkages [9–14,35]. If the spatial domain of a microstructure is discretized with equally sized square bins, such as the case with synthetic microstructures, the spatial correlation statistics is defined as

$$f_t^{np} = \frac{1}{S}\sum_{s=1}^{S} m_s^n m_{s+t}^p, \tag{1}$$

where f_t^{np} is the probability of finding a local state n and p in spatial bins separated by a vector t. The m_s^n is 1 if the local state n exists in the spatial bin marked s. Here, the local states n and p represent phases$n$ and p, respectively. Only the autocorrelation of the hard phase was computed for this study because a single two-point correlation captures all independent two-point correlation statistics [36]. In general, two-point correlation statistics qualifies as a good microstructure descriptor for many complex structures [37]. In fact, almost perfect recovery of the original microstructure is possible using phase recovery algorithm with the information contained within the statistics [38].

The correlation statistics comprises a very high dimensional data. Nevertheless, for most surrogate models and machine learning techniques, large dimensional features require a large amount data. Clearly, it is unfeasible to conduct sufficient number of simulations for 27 × 27 × 27 (19,683) features. Thus, dimensionality reduction is necessary. Several approaches such as principal component analysis (PCA) and multidimensional scaling (MDS) can be used for dimensionality reduction. The PCA can project data into low-dimensional space by using orthogonal transformation to form linearly uncorrelated variables. This transformation is carried out such that the first principal component has the largest variance, so one can select only the first few principal components to account for large variability in the data. This technique has been effective in several literatures establishing PSP linkages [9–14]. On the other hand, MDS only uses the distance matrix for dimension reduction and is specifically designed to preserve between-data distances as well as possible [39]. The main advantage of MDS over PCA is that MDS conducts nonlinear mapping of the data to a lower-dimensional space [40]. Fig. 2 demonstrates the performance of PCA and MDS in preserving pairwise Euclidean distances of data points when the number of dimensions is reduced to 5 and 8. One can see that while both methods perform well over large distances (Fig. 2(a)), MDS outperforms PCA in all cases for shorter distance ranges (Fig. 2(b)), which are very important because the Gaussian process regression (GPR) used in this study utilizes kernel method. The 8D MDS was used for low dimensional representation of microstructures.

2.4. Microstructure based simulations

The mechanical responses of synthetic microstructures were evaluated using 3D finite element method. A uniaxial tension was applied to the microstructures composed of 27 × 27 × 27 (19,683) cuboidal C3D8 elements with isotropic von Mises yield criterion using commercial ABAQUS 6.9 software. A simple constitutive equation was used to describe the flow stress (σ) of each phase

$$\sigma_i = k_i(1 + C_i\varepsilon_p)^{n_i}, \tag{2}$$

where ε_p, k_i, C_i, and n_i are the equivalent plastic strain and constitutive parameters, respectively. Table 1 shows the input parameters used for all FE simulations and Fig. 3 shows the resulting flow stress of each phase. All of the constitutive parameters were fixed so that microstructure is the only variable. A total of 35% elongation was imposed with periodic boundary condition. The average CPU time for FE simulations ranged from 15 to 46 min depending on the microstructure on a Dell Precision T7910 workstation (single 2.40 GHz CPU, 32 GB RAM).

2.5. Gaussian process regression

An observed value y that differ from a model $f(\mathbf{x})$ by an additive Gaussian noise with zero mean and a variance σ^2 can be represented as follows

$$y = f(\mathbf{x}) + \mathcal{N}(0, \sigma^2), \tag{3}$$

where $\mathbf{x}$ is the input vector of the regression function that is modeled as a Gaussian process [25]. Because the probability distribution over $f(\mathbf{x})$ follows the Gaussian distribution, $f(\mathbf{x})$ can be described by its mean $\mu(\boldsymbol{x})$ and covariance matrix $\mathbf{\Sigma}$ defined as

$$\mu(\mathbf{x}) = E(f(\mathbf{x})) \tag{4}$$

$$\mathbf{\Sigma} = \begin{bmatrix} E[(\mu(x_1)-f(x_1))(\mu(x_1)-f(x_1))] & \cdots & E[(\mu(x_1)-f(x_1))(\mu(x_n)-f(x_n))] \\ \vdots & \ddots & \vdots \\ E[(\mu(x_n)-f(x_n))(\mu(x_1)-f(x_1))] & \cdots & E[(\mu(x_n)-f(x_n))(\mu(x_n)-f(x_n))] \end{bmatrix} \tag{5}$$

where the operator $E()$ denotes the expected value of its argument and n represents the number of data. For simplicity, if the function has a mean μ and a covariance matrix $\mathbf{\Sigma}$, the joint distribution of a set of observed function values $\mathbf{f}$ and a set of test function values $\mathbf{f}_*$ with a mean of μ_*, according to Gaussian Process prior, can be given as [28]

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu \\ \mu_* \end{bmatrix}, \begin{bmatrix} \mathbf{\Sigma} + \sigma^2\mathbf{I} & \mathbf{\Sigma}_* \\ \mathbf{\Sigma}_*^{\mathrm{T}} & \mathbf{\Sigma}_{**} \end{bmatrix}\right), \tag{6}$$

where $\mathbf{\Sigma}_*$ and $\mathbf{\Sigma}_{**}$ correspond to the covariance matrix of the observed and the test values and the covariance matrix for the test values, respectively. This joint distribution represents that $\mathbf{f}$ is modeled by μ and $\mathbf{\Sigma} + \sigma^2\mathbf{I}$ while the relationship between $\mathbf{f}$ and $\mathbf{f}_*$ are modeled through $\mathbf{\Sigma}_*$. Then, the posterior distribution, or the distribution of $\mathbf{f}_*$ given $\mathbf{f}$, for a specific set of test cases can be derived as

$$\mathbf{f}_*|\mathbf{f} \sim \mathcal{N}(\mu_* + \mathbf{\Sigma}_*^{\mathrm{T}}(\mathbf{\Sigma} + \sigma^2)^{-1}(\mathbf{f}-\mu), \mathbf{\Sigma}_{**}-\mathbf{\Sigma}_*^{\mathrm{T}}(\mathbf{\Sigma} + \sigma^2)^{-1}\mathbf{\Sigma}_*). \tag{7}$$

If one replaces $\mathbf{\Sigma}$, $\mathbf{\Sigma}_*$, and $\mathbf{\Sigma}_{**}$ with a covariance function $k(\mathbf{x}, \mathbf{x}')$, the posterior will then only depend on observed values, the covariance function, and the test input vector. Generally, the covariance function will be defined using hyperparameters $\Theta = (\theta_1, ..., \theta_{\boldsymbol{i}})$, which can be determined by maximizing the log marginal likelihood (L) in terms of the hyperparameters [28]. L is defined as

$$L = \log p(\mathbf{y}|\mathbf{x}, \Theta) = -\frac{1}{2}\log(\det(k(\mathbf{x}, \mathbf{x}') + \sigma^2\mathbf{I}))-\frac{1}{2}\mathbf{y}^{\mathrm{T}}[k(\mathbf{x}, \mathbf{x}') + \sigma^2\mathbf{I}]^{-1}\mathbf{y} -\frac{n}{2}\log 2\pi \tag{8}$$

where $p(\mathbf{y}|\mathbf{x}, \Theta)$ represents the likelihood of $\mathbf{y}$ given $(\mathbf{x}, \Theta)$.

One of the advantages of GPR model is that one can obtain the uncertainty of the model prediction in terms of the predicted variance K as [28]

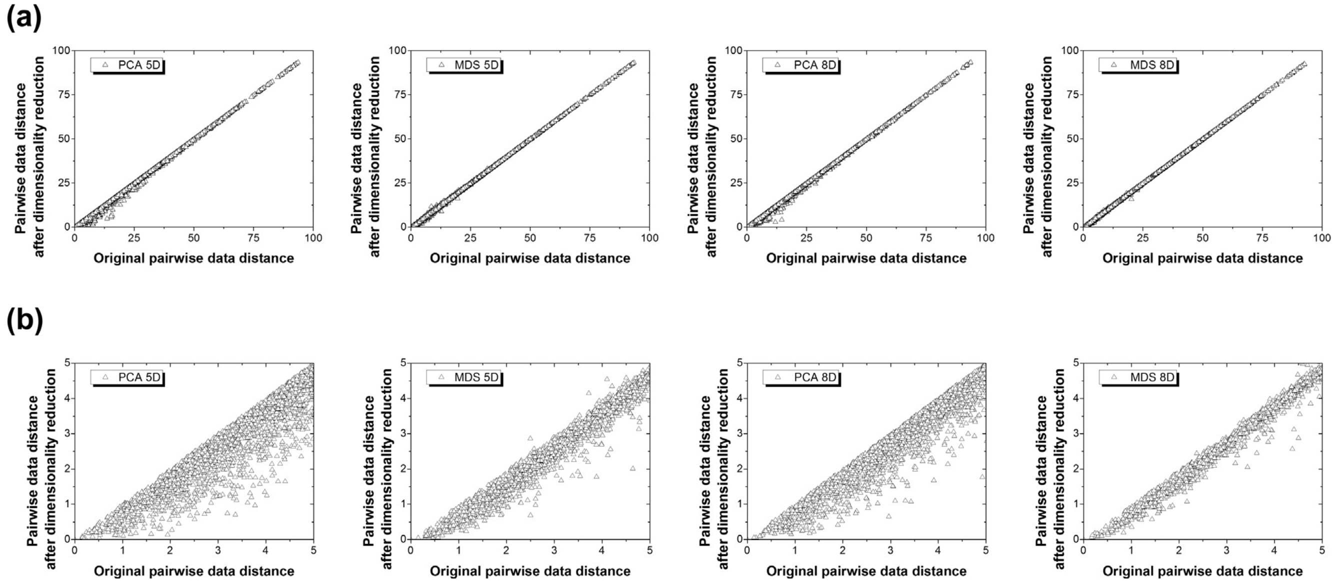

Fig. 2. Performance of multi-dimensional scaling principal component analysis in terms of pairwise data distance preservation (a) over all distance range and (b) over small distance range.

Table 1
Constitutive parameters used for all finite element simulations.

	k_i	C_i	n_i
Soft phase	300	55	0.33
Hard phase	1700	100	0.08

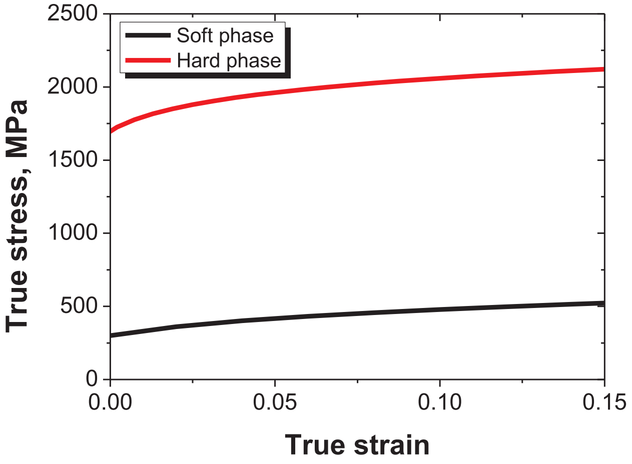

Fig. 3. Flow stress of each phase.

$$K = k(\mathbf{x}_*, \mathbf{x}_*) - k(\mathbf{x}_*, \mathbf{x})(k(\mathbf{x}, \mathbf{x}) + \sigma^2)^{-1} k(\mathbf{x}, \mathbf{x}_*). \tag{9}$$

In this work, a covariance function based on radial basis function (RBF) was selected due to the simplicity and flexibility of the RBF. The covariance function is defined as

$$k(\mathbf{x_i}, \mathbf{x}) = \sigma^2 \exp\left(-\theta \frac{\|\mathbf{x_i} - \mathbf{x}^2\|}{l^2}\right), \tag{10}$$

where θ is the hyperparameter that needs to be optimized and l is the length scale. Unlike the hyperparameter, the length scale should be predefined. In this study, a semivariogram was used to determine the length scale. Semivariogram depicts how semivariance (γ) changes as the distance between observed point changes [41]. This can be measured using initially sampled data by

$$\gamma(h) = \frac{1}{2N(h)} \sum_{(i,j)\in N(h)} |f_i - f_j|^2, \tag{11}$$

where h, $N(h)$, and $|f_i - f_j|$ represent the pairwise data distance, number of pair of data separated by h, difference between observed function values whose pairwise data distance is separated by h.

2.6. Efficient global optimization

A key advantage of Gaussian process regression (GPR) is that GPR models are apt for optimization purposes [28,42,43], in particular with expensive black-box simulations using the EGO method [28]. The optimization procedure using the EGO method utilizes expected improvement (EI) function defined as

$$EI = \begin{cases} (f_{min} - f^*_{min})\Phi\left(\frac{f_{min} - f^*_{min}}{\sigma_E}\right) + \sigma_E \phi\left(\frac{f_{min} - f^*_{min}}{\sigma_E}\right) & if \ \sigma_E > 0 \\ 0 & if \ \sigma_E = 0 \end{cases}, \tag{12}$$

where f_{min}, f^*_{min}, Φ, and ϕ are minimum of observed function values ($\mathbf{f}$), minimum of test function values ($\mathbf{f}_*$), cumulative distribution function, and probability density function, respectively. By introducing both possible minimum value f^*_{min} and its predicted variance σ_E, the EI balances minimization with uncertainty. That is, the EI function will be large where f^*_{min} is likely to be smaller than the observed minimum and/or where there is high uncertainty.

3. Results and discussion

3.1. Microstructure based simulation results

Microstructure based simulations were conducted over 1100 different synthetic microstructures to calculate the uniform elongation (U.El), ultimate tensile strength (UTS), and strain localization index (X_1) defined in soft phase of each microstructure. With ε_i representing the equivalent plastic strain of phase i, the strain localization index of the soft phase is defined as

$$X_1 = \frac{\varepsilon_1}{\varepsilon_{bulk}}. \tag{13}$$

For uniaxial tension, U.El and UTS can be found using Considère criterion defined as

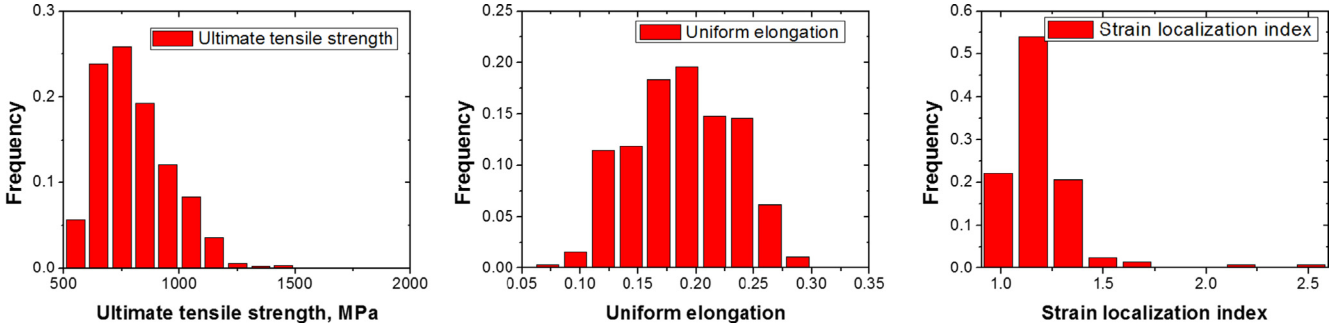

Fig. 4. Distribution of mechanical properties of microstructures within the microstructure database.

$$\frac{\partial \sigma_{bulk}}{\partial \varepsilon_{bulk}} = \sigma_{bulk}, \tag{14}$$

where σ_{bulk} is the bulk flow stress defined as von Mises stress calculated with average stress components over the microstructure volume element. Likewise, ε_{bulk} is defined as the von Mises strain calculated with average strain components over the microstructure volume element. The stress and strain that satisfy Eq. (14) will be the UTS and U.El, respectively. Distributions of U.El, UTS, and X_1 show that the microstructure dataset consists of different properties that stem from different microstructures (Fig. 4). Clearly, the constructed database consists of diverse microstructures. Fig. 5 represents three groups of microstructures that exhibit high U.El, UTS, or low X_1 values. Qualitatively, microstructures with high U.El and UTS can be characterized by high volume fraction of hard phase. Likewise, microstructures with high UTS exhibit low U.El and vice versa, indicating a direct correlation among UTS, U.El, and volume fraction of the hard phase. On the other hand, microstructures with low X_1 do not seem to share intuitively recognizable microstructural features.

3.2. Constructing structure-property linkages

A semivariogram was generated using the results of full-field simulations on 50 microstructures randomly selected from the microstructure database to quickly assess the length scale of the kernel function. According to the semivariogram shown in Fig. 6, the normalized semivariances for UTS, U.El, and X_1 level out within pairwise distances from 4.0 to 6.0. Therefore, a length scale of 4.0 was used for all UTS, U.El, and X_1. To demonstrate the efficiency the proposed approach in constructing SP linkages, only 10 of the 50 microstructures used to construct the semivariograms were selected for initial sampling data. All output variables used for GPR were normalized to have zero mean. Predictive performance of the GPR model for was validated with more than 1000 microstructures that were not used in training the model. Separate models for UTS, U.El, and X_1 were built to presume uncorrelated outputs. The accuracy of the surrogate model prediction was evaluated using percent relative error defined as

$$100 \times \left|\frac{\text{Exact value} - \text{Approximation}}{\text{Exact value}}\right| \tag{15}$$

where exact value is the finite element simulated result and approximation is the surrogate model prediction.

In the case of UTS, the predictive performance of the GPR model was relatively low when 10 microstructures were used to train the model. The maximum predicted variance, or maximum uncertainty, was 0.2. Then, the next full-field simulation was conducted with the microstructure with the highest uncertainty. The predictive performance of the GPR model after 10 to 40 full-field simulations by sampling data with highest predicted variance gradually increased with maximum relative error decreasing from 43% to 5.4% (Fig. 7(a)). The resulting predictions using the GPR model for 1000 microstructures are shown in Fig. 7(b). The highest and average relative errors after only 40 simulations were 5.4% and 0.5%, respectively. In the case of U.El, a total of 61 full-field simulations were required to satisfy the stopping criteria. The predictive performance of GPR model for U.El after 10 to 61 full-field simulations is shown in Fig. 7(c). Initially, the maximum predicted variance as well as maximum relative error decreased drastically with increasing number of full-field simulation. Afterwards, maximum predicted variance and maximum relative error slowly

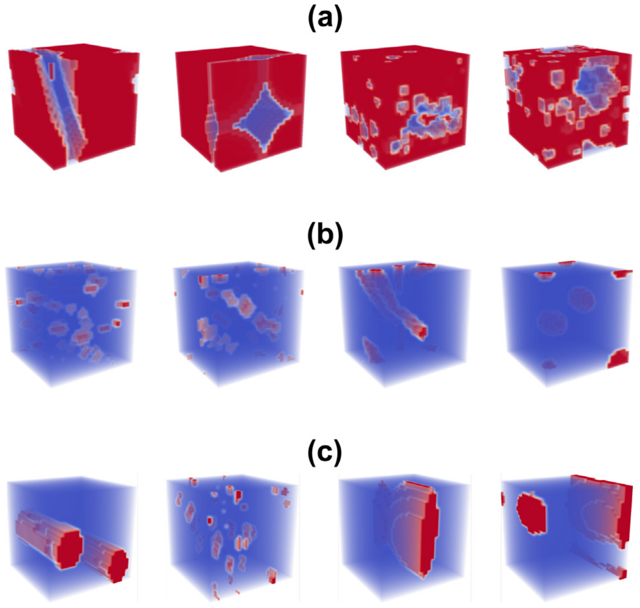

Fig. 5. Microstructures that exhibit very high (a) ultimate tensile strength, (b) uniform elongation, and low (c) strain localization index of soft phase.

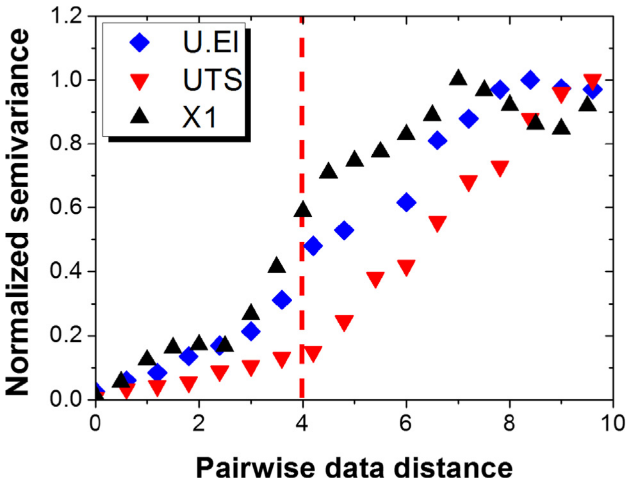

Fig. 6. Normalized semivariogram for UTS, U.El, and X1.

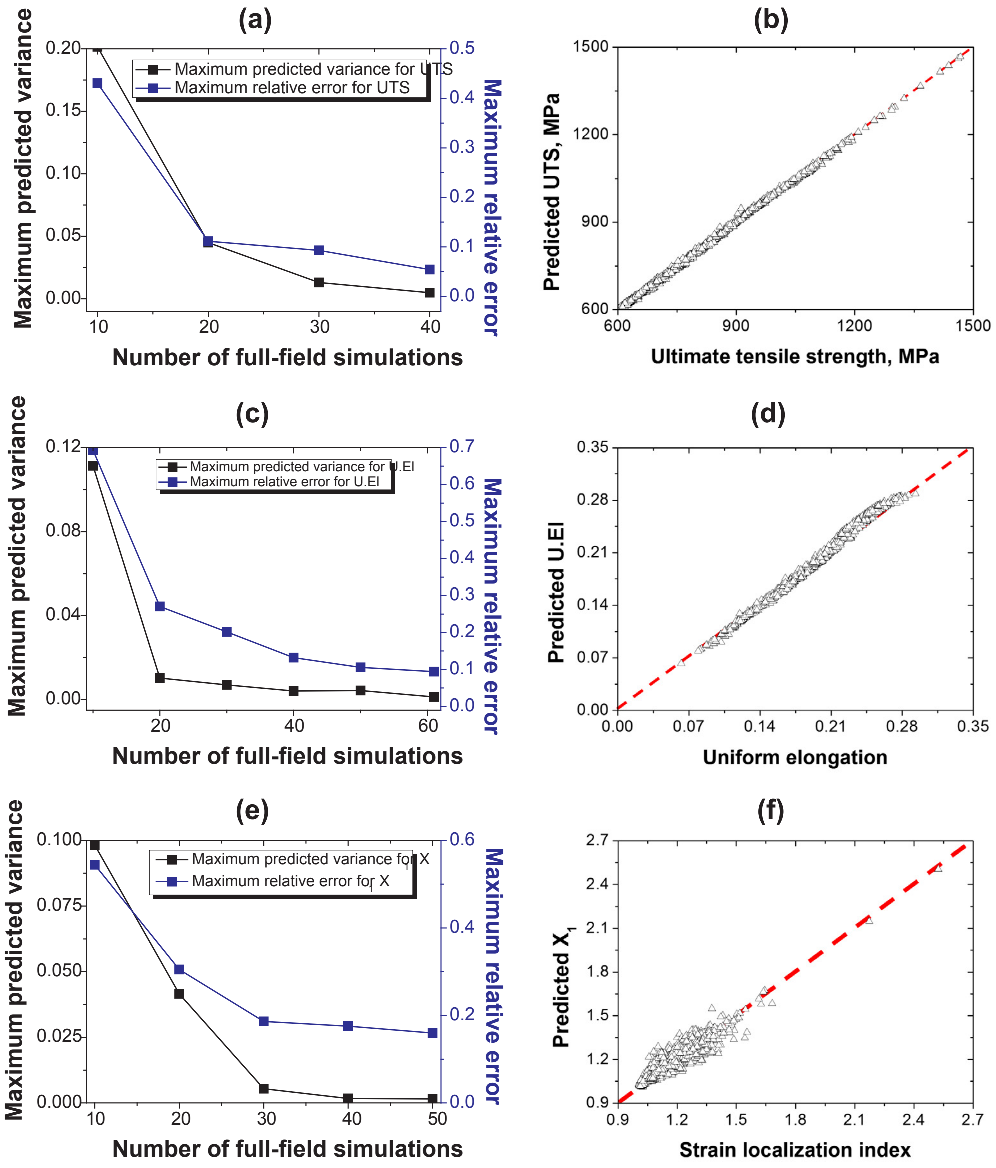

Fig. 7. Maximum predicted variance and maximum relative error (a, c, e) with respect to the number of simulations, and model predictions (b, d, f) for UTS (a, b), U.El (c, d), and X1 (e, f).

decreased to 0.0013 and 8.4%, respectively. The resulting predictions using the GPR model for 1000 microstructures are shown in Fig. 7(d). The highest and average relative errors after only 61 simulations were 8.4% and 2.4%, respectively. To satisfy the stopping criteria for X_1, 50 full-field simulations were necessary. Both maximum predicted variance and maximum relative error decreased with increasing number of full-field simulations (Fig. 7(e)). The maximum and average relative errors were 15.9% and 3.3% after 50 simulations, respectively. The resulting predictions using the GPR model for 1000 microstructures are shown in Fig. 7(f).

Because GPR training depends on initially sampled data, the procedures for training the models for all UTS, U.El, and X_1 were repeated four additional times using the randomly selected full-field simulations used to form the semivariogram. The number of full-field simulations necessary to reach the prescribed stopping criteria and the resulting maximum relative error are shown in Table 2. One can see from Table 2 that the simulations necessary to train GPR models for UTS, U.El, and X_1 fluctuates around 35, 50, and 47, respectively. All GPR models share a general trend in decrease in maximum predicted variance and maximum relative error with increase in the number of full-field simulations. The GPR models for UTS, U.El, and X_1 exhibited prediction accuracies above 80% over 1000 microstructures with only 30 to 60 full-field simulations. In particular, the models for UTS and U.El exhibited prediction accuracies generally above 90% with minor data spread. Nevertheless, the GPR model for X_1 exhibited a relatively lower prediction accuracy of 80% to 85% with data spread wider than those of the other models. The high performance of the models for UTS and U.El is attributed to the fact that the volume fraction of hard, or soft, phase is strongly correlated with these properties (Fig. 8(a) and (b)). The strain localization index, however, does not retain such high

Table 2
The number of simulations necessary for GPR training and the maximum relative error of the trained model.

	Number of simulations (Average/Min/Max)	Maximum relative error (Average/Min/Max)
UTS	35/27/40	8.1%/5.4%/15%
U.El	51/40/61	10.8%/7.7%/17.1%
X_1	47/43/53	17%/15.9%/20%

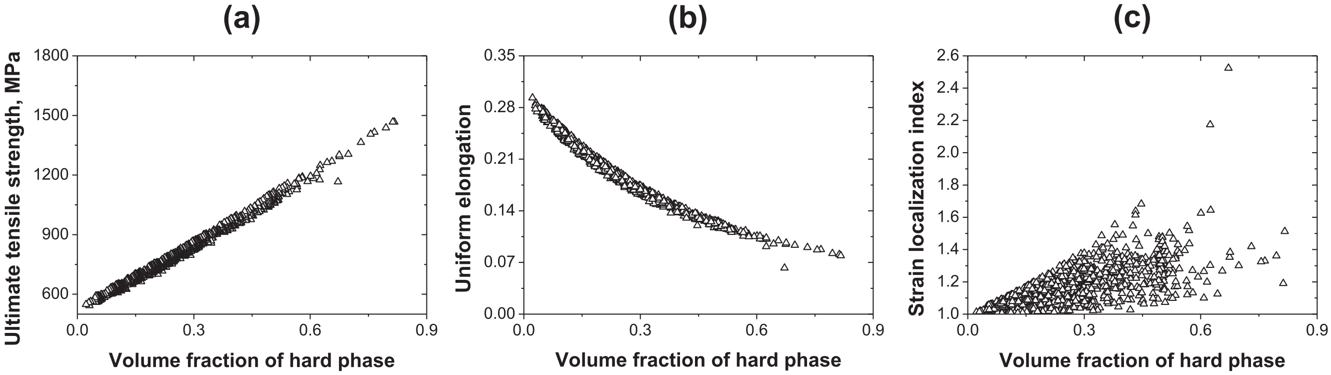

Fig. 8. Effect of volume fraction of hard phase on (a) UTS, (b) U.El, and (c) X1.

correlation with volume fraction of hard or soft phase (Fig. 8(c)). However, it is worth noting that the predictive performance of GPR models is quantified with maximum relative error. In the case of X_1, even though the maximum relative error was between 15% to 20%, only 3% to 4% of the entire dataset exhibit relative error above 10%.

3.3. Finding optimal microstructures

The EGO method is a Bayesian optimization technique that utilizes EI algorithm to maximize an unknown black-box objective function. Unlike GPR, which samples the point offering the greatest amount of information, the EI algorithm samples point where expected improvement (Eq. (12)) is high. We demonstrated that GPR model itself can achieve a rather high accuracy with generally less than 50 full-field simulations, meaning that optimal microstructures can be found within 50 simulations even without implementing EI algorithm. Therefore, in this particular set of problems, if the implementation of EI algorithm requires more than 50 full-field simulations to find an optimal microstructure, there is no need to use the algorithm.

After 10 randomly sampled data, the maximum EI was 0.15 with predicted maximum UTS of 941 MPa. Surprisingly, the stopping criteria was met after only 8 additional full-field simulations. The predicted maximum UTS was 1467 MPa, which is identical to the actual maximum, and the microstructure with the predicted maximum UTS is shown in Fig. 9(a), which is identical to the one shown in Fig. 5. That is, only 18 full-field simulations were necessary to find the microstructure with the highest UTS. Of course, the total number of full-field simulations seems to fluctuate depending on the initially sampled data. In some cases, more than 20 full-field simulations were necessary, but the number of simulations necessary to accurately predict maximum UTS was generally less than 20. In the case of U.El, the total number of simulations necessary ranged from 20 to 40. The predicted maximum U.El values were generally around 0.24 to 0.29, which is similar to the actual maximum, but the predicted optimal microstructures, which is shown in Fig. 9(b), were not always identical to the actual optimal microstructure that is shown in Fig. 5. This is primarily due to the fact that the GPR model itself exhibits highest relative error near the high U.El range as illustrated in Fig. 10. However, even when the predicted optimal microstructure was not the actual optimal microstructure, the predicted microstructure was always one of the microstructures with U.El values within top 1%. In the case of lowest X_1, 22 to 32 full-field simulations, depending on the initially sampled data, were necessary to satisfy the stopping criteria. The predictive minimum was generally within the range of 1.007 to 1.01, which is in good agreement with the actual minimum. The predicted optimal microstructure (Fig. 9(c)) was not identical to the actual optimal microstructure because there were ten microstructures with $X_1 < 1.01$. Nonetheless, the predicted microstructure was always one of the ten microstructures.

In general, implementation of EI algorithm was effective in finding the optimal microstructure within a large database. The optimal microstructures to maximize UTS or U.El, or to minimize X_1 can be found with only 18 to 40 full-field simulations. Additionally, even when predicted optimal microstructure was not the actual optimal microstructure, the predicted microstructure was still one of the most desirable microstructures in terms of targeted property. Therefore, if one is solely interested in finding optimal microstructures, application of EI algorithm can be used instead of uncertainty driven sampling.

3.4. Efficiency of the proposed approach

One of the main advantages the proposed approach offer is efficiency that originates from robust surrogate modelling as well as an efficient sampling scheme The trained surrogate model takes 0.2 s in making 1000 predictions using a personal desktop (single 3.40 GHz CPU, 16.0 GB RAM), which is clearly much faster than FE simulations. Furthermore, the uncertainty driven sampling scheme enables accurate surrogate model to be trained with only 34 to 61full-field simulations depending on the property to predict (Table 2). If one is to randomly sample the training data, the maximum relative error values in predicting X_1 using surrogate models trained with 100, 200, and 400 data are 40%, 34%, 23%, respectively. Thus, implementing uncertainty driven sampling scheme reduces the number of simulations necessary to

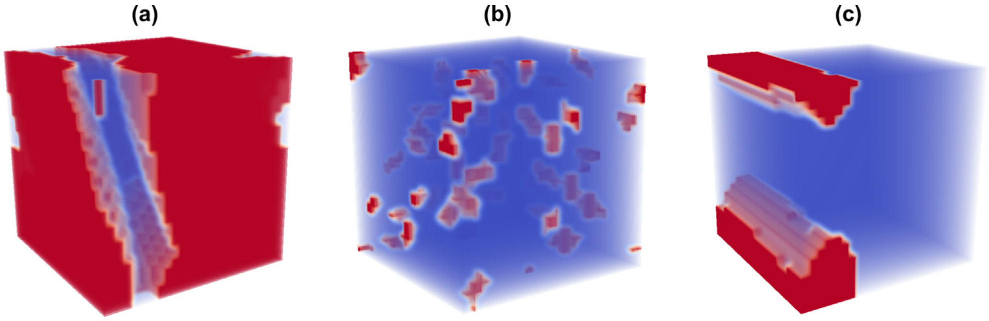

Fig. 9. Predicted optimal microstructures that (a) maximize UTS, (b) maximize U.El, and minimize (c) X1.

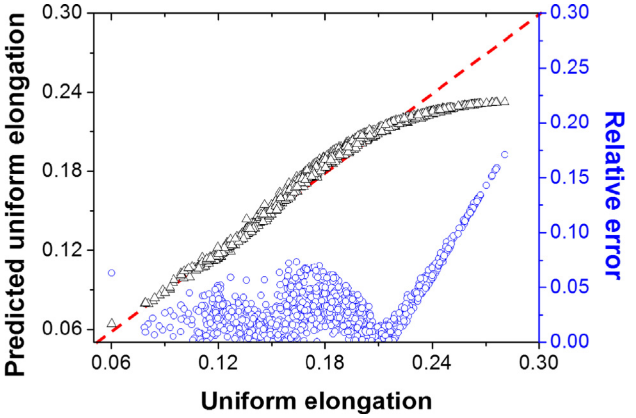

Fig. 10. Accuracy of the predicted U.El when maximum relative error reaches up to 17%.

achieve high accuracy down from 400 or more to less than 61.

3.5. Possible strategies for improvement

One of the most effective methods to improve the efficiency of GPR is to reduce the number of input features while preserving the pairwise data distance as well as possible. Lubbers et al. [39] demonstrated that the combined use of Gram matrices from filtered images of convolutional neural network (CNN) and MDS can perform better than the combined use of two-point correlation and MDS. On the other hand, when not all input features are correlated with output features, feature selection prior to GPR can improve the performance of GPR models. In particular, for mechanical properties such as UTS and U.El that are highly correlated with input features, feature selection can improve the performance of GPR models. While there are a number of approach possible for feature selection [44–46], simple filter method seems to be sufficient for cases like UTS and U.El [45]. Using the simplest filter method based on Pearson correlation coefficient, one can easily determine that a single feature prevails over all other features. If only the features with very high correlation coefficient are chosen as the input variable, the number of training data required for the GPR models for UTS and U.El can be drastically reduced. Nonetheless, even the simplest filter method such as Pearson correlation coefficient requires data. Thus, one needs to take into account the number of data required to obtain Pearson correlation coefficients. Increase in efficiency is only possible if the number of data to obtain Pearson correlation coefficients is less than the decrease in the number of full-field simulations necessary to train GPR by reducing the number of input variables.

4. Conclusions

The proposed framework provides a very efficient means of computationally aided materials design by training GPR models with low dimensional representation of microstructures as inputs and full-field simulations as outputs. Using the GPR with uncertainty driven sampling scheme, we established accurate SP linkages over a microstructure database composed of over 1100 synthetic microstructures by evaluating only a small fraction of the entire microstructure database. Moreover, using the GPR with EI based sampling scheme, we were able to find the optimal microstructure that maximizes a specified property with even fewer number of full-field evaluations. The framework is applicable to a wide range of materials related optimization problems granted an accurate microstructure based modelling is available. Furthermore, this framework can incorporate experimental data along with simulation data. The framework relies on accurate physical model to produce reliable data for training. This means that if full-field simulations are conducted with accurate physical model and model parameters are well-calibrated to reflect experimental data, then the proposed framework can easily incorporate experimental data. A straightforward application of the proposed approach is in finding optimal microstructure of dual-phase (DP) steel that exhibit maximum yield strength or ultimate tensile strength, or minimum yield ratio or strain localization index. One can start by analyzing existing experimental data on microstructures to obtain variable microstructural features such as size, shape, and orientation distributions of each phase. Then, a compact database can be built using synthetic microstructures generated by varying the experimentally obtained microstructural features. Afterwards, physical constitutive model and its model parameters are calibrated to reflect experimental result on property of interest so that microstructure-based simulations can produce reliable data. Finally, by applying GPR with low-dimensional representation of microstructure as input, one can find optimal microstructure for targeted property.

Data availability

The raw/processed data required to reproduce these findings cannot be shared at this time as the data also forms part of an ongoing study.

Acknowledgment

This study was supported by POSCO (2017Y054) and Brain Korea 21 PLUS project for Center for Creative Industrial Materials (F16SN25D1706).

Author contribution

Jaimyun Jung and Hyung Keun Park contributed to the implementation of the computer code and supporting algorithms. Jaimyun Jung and Hyoung Seop Kim have contributed to development of the methodology and preparation of the manuscript. Jaimyun Jung, Jae Ik Yoon, Jin You Kim, and Hyoung Seop Kim have contributed to conceptualization.

References

[1] A. Mikdam, A. Makradi, S. Ahzi, H. Garmestani, D.S. Li, Y. Remond, Effective conductivity in isotropic heterogeneous media using a strong-contrast statistical continuum theory, J. Mech. Phys. Solids 57 (2009) 76–86.

[2] E. Kröner, Bounds for effective elastic moduli of disordered materials, J. Mech. Phys. Solid. 25 (1977) 137–155.

[3] W.F. Brown, Solid mixture permittivities, J. Chem. Phys. 23 (1955) 1514–1517.

[4] B. Raju, S.R. Hiremath, D.R. Mahapatra, A review of micromechanics based models for effective elastic properties of reinforced polymer matrix composites, Compos. Struct. 204 (2018) 607–619.

[5] L.F.A. Bernardo, A.P.B.M. Amaro, D.G. Pinto, S.M.R. Lopes, Modeling and simulation techniques for polymer nanoparticle composites – a review, Comput. Mater. Sci. 118 (2016) 32–46.

[6] A. Abdul-Latif, A. Kerkour-El Miad, R. Baleh, H. Garmestani, Modeling the mechanical behavior of heterogeneous ultrafine grained polycrystalline and nanocrystalline FCC metals, Mech. Mater. 126 (2018) 1–12.

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

[8] S. Torquato, Effective stiffness tensor of composite media – I. Exact series expansions, J. Mech. Phys. Solids 45 (1997) 1421–1448.

[9] M.I. Latypov, S.R. Kalidindi, Data-driven reduced order models for effective yield strength and partitioning of strain in multiphase steels, J. Comput. Phys. 346 (2017) 242–261.

[10] Y.C. Yabansu, P. Steinmetz, J. Hötzer, S.R. Kalidindi, B. Nestler, Extraction of reduced-order process-structure linkages from phase-field simulations, Acta Mater. 124 (2017) 182–194.

[11] N.H. Paulson, M.W. Priddy, D.L. McDowell, S.R. Kalidindi, Reduced-order structure-property linkages for polycrystalline microstructures based on 2-point statistics, Acta Mater. 129 (2017) 428–438.

[12] A. Khosravani, A. Cecen, S.R. Kalidindi, Development of high throughput assays for establishing process-structure-property linkages in multiphase polycrystalline metals: application to dual-phase steels, Acta Mater. 123 (2017) 55–69.

[13] A. Iskakov, Y.C. Yabansu, S. Rajagopalan, A. Kapustina, S.R. Kalidindi, Application of spherical indentation and the materials knowledge system framework to establishing microstructure-yield strength linkages from carbon steel scoops excised from

high-temperature exposed components, Acta Mater. 144 (2017) 758–767.
[14] Y.C. Yabansu, D.K. Patel, S.R. Kalidindi, Calibrated localization relationships for elastic response of polycrystalline aggregates, Acta Mater. 81 (2014) 151–160.
[15] H. Zhang, M. Diehl, F. Roters, D. Raabe, A virtual laboratory using high resolution crystal plasticity simulations to determine the initial yield surface for sheet metal forming operations, Int. J. Plast. 80 (2016) 111–138.
[16] H.K. Park, J. Jung, H.S. Kim, Three-dimensional microstructure modeling of particulate composite using statistical synthetic structure and its thermos-mechanical finite element analysis, Comput. Mater. Sci. 126 (2017) 265–271.
[17] E.-Y. Kim, S.I. Kim, S.-H. Choi, Effect of strength coefficient of bainite on micromechanical deformation and failure behaviors of hot-rolled 590FB steel during uniaxial tension, Korean, J. Met. Mater. 54 (2016) 808–816.
[18] J. Jung, J.I. Yoon, J.G. Kim, M.I. Latypov, J.Y. Kim, H.S. Kim, Continuum understanding of twin formation near grain boundaries of FCC metals with low stacking fault energy, npj Comput. Mater. 3 (2017) 21.
[19] S.L. Wong, M. Madivala, U. Prahl, F. Roters, D. Raabe, A crystal plasticity model for twinning- and transformation-induced plasticity, Acta Mater. 118 (2016) 140–151.
[20] F. Maresca, V.G. Kouznetsova, M.G.D. Greers, Deformation behaviour of lath martensite in multi-phase steels, Scr. Mater. 110 (2016) 74–77.
[21] J. Jung, J.I. Yoon, J.H. Moon, H.K. Park, H.S. Kim, Effect of coarse precipitates on surface roughening of an FCC polycrystalline material using crystal plasticity, Comput. Mater. Sci. 126 (2017) 121–131.
[22] S. Keshavarz, S. Ghosh, A.C.E. Reid, S.A. Langer, A non-Schmid crystal plasticity finite element approach to multi-scale modeling of nickel-based superalloys, Acta Mater. 114 (2016) 106–115.
[23] F. Farukh, L.G. Zhao, R. Jiang, P. Reed, D. Proprentner, B.A. Shollock, Realistic microstructure-based modelling of cyclic deformation and crack growth using crystal plasticity, Comput. Mater. Sci. 111 (2016) 395–405.
[24] T. Chugh, N. Chakraborti, K. Sindhya, Y. Jin, A data-driven surrogate-assisted evolutionary algorithm applied to a many-objective blast furnace optimization problem, Mater. Manuf. Process. 32 (2017) 1172–1178.
[25] A.P. Bartók, M.C. Payne, R. Kondor, G. Csányi, Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons, Phys. Rev. Lett. 104 (2010) 136403.
[26] A. Chowdhury, E. Kautz, B. Yener, D. Lewis, Image driven machine learning methods for microstructure recognition, Comput. Mater. Sci. 123 (2016) 176–187.
[27] M. Rupp, A. Tkatchenko, K.R. Müller, O.A. von Lilienfeld, Fast and accurate modeling of molecular atomization energies with machine learning, Phys. Rev. Lett. 108 (2012) 058301.
[28] D.R. Jones, M. Schonlau, W.J. Welch, Efficient global optimization of expensive black-box functions, J. Global Optim. 13 (1998) 455–492.
[29] V. Nguyen, S. Gupta, S. Rana, C. Li, S. Venkatesh, Regret for expected improvement over the best-observed value and stopping condition, PMLR 77 (2017) 279–294.
[30] J. Zhang, A.A. Taflanidis, J.C. Medina, Sequential approximate optimization for design under uncertainty problems utilizing Kriging metamodeling in augmented input space, Comput. Methods Appl. Mech. Eng. 315 (2017) 369–395.
[31] D. Drignei, An estimation algorithm for fast kriging surrogates of computer models with unstructured multiple outputs, Comput. Methods Appl. Mech. Eng. 321 (2017) 35–45.
[32] V. Picheny, D. Ginsbourger, Noisy kriging-based optimization methods: a unified implementation within the DiceOptim package, Comput. Stat. Data Anal. 71 (2014) 1035–1053.
[33] A. Matthews, M. van der Wilk, T. Nickson, K. Fujii, A. Boukouvalas, P. León-Villagrá, Z. Ghahramani, J. Hensman, GPflow: a Gaussian process library using tensor flow, J. Mach. Learn. Res. 18 (2017) 1–6.
[34] M.A. Groeber, M.A. Jackson, DREAM.3D: a digital representation environment for the analysis of microstructure in 3D, Integr. Mater. Manuf. Innov. 3 (2014) 1–17.
[35] A. Choudhury, Y.C. Yabansu, S.R. Kalidindi, A. Dennstedt, Quantification and classification of microstructures in ternary eutectic alloys using 2-point spatial correlations and principal component analyses, Acta Mater. 110 (2016) 131–141.
[36] A.M. Gokhale, A. Tewari, H. Garmestani, Constraints on microstructural two-point correlation functions, Scr. Mater. 53 (2005) 989–993.
[37] M. Yang, A. Nagarajan, B. Liang, S. Soghrati, New algorithms for virtual reconstruction of heterogeneous microstructures, Comput. Methods Appl. Mech. Eng. 338 (2018) 275–298.
[38] 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.
[39] N. Lubbers, T. Lookman, K. Barros, Inferring low-dimensional microstructure representations using convolutional neural networks, Phys. Rev. E 96 (2017) 052111.
[40] V.S. Sumithra, S. Surendran, A review of various linear and non linear dimensionality reduction techniques, IJCSIT 6 (2015) 2354–2360.
[41] P.M. Atkinson, P. Lewis, Geostatistical classification for remote sensing: an introduction, Comput. Geosci. 26 (2000) 361–371.
[42] L. Lebensztajin, C.A.R. Marretto, M.C. Costa, J.-L. Coulomb, Kriging: a useful tool for electromagnetic device optimization, IEEE Trans. Magn. 40 (2004) 1196–1199.
[43] Z. Hu, D. Ao, S. Mahadevan, Calibration experimental design considering field response and model uncertainty, Comput. Methods Appl. Mech. Eng. 318 (2017) 92–119.
[44] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, J.M. Benítez, F. Herrera, A review of microarray datasets and applied feature selection methods, Inf. Sci. 282 (2014) 111–135.
[45] G. Chandrashekar, F. Sahin, A survey on feature selection methods, Comput. Electr. Eng. 40 (2014) 16–28.
[46] Q. Lu, X. Li, Y. Dong, Structure persevering unsupervised feature selection, Neurocomputing 301 (2018) 36–45.