Hu *et al. J Mater Inf* 2022;2:5
DOI: 10.20517/jmi.2022.05

Journal of materials Informatics

Research Article

Open Access

Structure-property modeling scheme based on optimized microstructural information by two-point statistics and principal component analysis

Xiaobing Hu, Jiajun Zhao, Yiming Chen, Yujian Wang, Junjie Li, Qingfeng Wu, Zhijun Wang, Jincheng Wang

State Key Laboratory of Solidification Processing, Northwestern Polytechnical University, Xi'an 710072, China.

Correspondence to: Prof. Jincheng Wang, State Key Laboratory of Solidification Processing, Northwestern Polytechnical University, 127 West Youyi Road, Xi'an 710072, China. E-mail: jchwang@nwpu.edu.cn; Prof. Junjie Li, State Key Laboratory of Solidification Processing, Northwestern Polytechnical University, 127 West Youyi Road, Xi'an 710072, China. E-mail: lijunjie@nwpu.edu.cn

How to cite this article: Hu X, Zhao J, Chen Y, Wang Y, Li J, Wu Q, Wang Z, Wang J. Structure-property modeling scheme based on optimized microstructural information by two-point statistics and principal component analysis. *J Mater Inf* 2022;2:5. http://dx.doi.org/10.20517/jmi.2022.05

Received: 24 Mar 2022 **First Decision:** 20 Apr 2022 **Revised:** 3 May 2022 **Accepted:** 18 May 2022 **Published:** 27 May 2022

Academic Editor: Xingjun Liu **Copy Editor:** Tiantian Shi **Production Editor:** Tiantian Shi

Abstract

Construction of the structure-property (SP) relationship is an important tenet during materials development. Optimizing microstructural information is a necessary and challenging task in understanding and improving this linkage. To solve the problem that the experimental microstructures with a small size usually fail to represent the entire sample structure, a data-driven scheme integrating two-point statistics, principal component analysis, and machine learning was developed to reasonably construct a representative volume element (RVE) set from the small microstructures and extract optimized structural information. Based on the elaborate quantitative metrics and method, this kind of RVE set was successfully constructed on an experimental microstructure dataset of ferrite heat-resistant steels. Moreover, to remove redundant information included in two-point statistics, the critical threshold of the tolerance factor related to the coherence length in microstructures was determined to be 0.005. An accurate SP linkage was finally established (mean absolute error < 6.28MPa for yield strength). This scheme was further validated on two other simulated and experimental datasets, which proved that it can offer scientific nature, reliability, and universality compared to traditional strategies. This scheme has a bright application prospect in microstructure classification, property prediction, and alloy design.

Keywords: Microstructural information, structure-property linkage, two-point statistics, machine learning

© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

INTRODUCTION

A great acceleration in target prediction and alloy design can be realized through materials informatics including multitudinous advanced data-driven technologies and theories related to materials science, which has received extensive attention in recent years[1–3]. Establishing process-structure-property (PSP) linkages represents a recognized core task for achieving this ambitious goal. Indeed, significant efforts have been made to pursue such an accurate and universal linkage[3–5].

Traditionally, materials development is largely completed by a mix of Edisonian approaches and serendipity, which can extract experiential process-property (PP) relationships from existing experimental data, and then studies to understand and explain the dominant mechanism leading to the expectations or serendipity through investigating microstructural features[6]. Using an informatics strategy that is different from the traditional experiment methods simply guided by physical metallurgy knowledges, we previously demonstrated the improvement effect of microstructural information in predicting the hardness of austenite steels by comparing PP and PSP linkages[7]. Similarly, Molkeri *et al.* proposed a novel microstructure-aware framework for materials design and rigorously confirmed the importance of microstructure information in alloy design[8]. One can find that the focus of attention on microstructure has gradually shifted from providing scientific explanations to practically promoting the forward and reverse process of PSP. Therefore, it is necessary to quantificationally extract microstructure information[9–13]. Generally, some physical parameters based on statistical average (phase fraction, grain size, *etc.*) are used to simply characterize the microstructural features of materials, which nevertheless ignores the correlation and heterogeneity of these features in terms of spatial distribution. In addition, considering the entire discrete and highly nonlinear micrograph data as partial input of a PSP linkage, one may be at risk of dimensional disaster due to the inapplicability of some common dimension reduction algorithms[14]. Therefore, a challenge to be addressed is how to extract sufficient and effective microstructure information including correlation and heterogeneity of spatial features in a quantitative and low-dimensional manner.

Some admirable efforts have been made to overcome this challenge. Sangid *et al.* used crystal plasticity simulations to identify the stress concentration around pores of various sizes and quantify the pore with the smallest size that results in a debit in the fatigue performance of IN718 alloy[15]. Zinovieva *et al.* proposed a multi-physics methodology combining physically based cellular automata to simulate the grain structure evolution[16]. They successfully uncovered the effects of scanning pattern on the microstructure and elastic properties of 316L austenitic stainless steel prepared by powder bed-based additive manufacturing. It is noted that, although the two works mentioned above used advanced simulation methods to investigate the PSP relationship, the high computational cost and high-dimensional data analysis process were not completely avoided. Popova *et al.* developed a data-driven workflow and applied it to a set of synthetic AM microstructures obtained using the Potts-kinetic Monte Carlo (kMC) approach[17]. They finally correlated process parameters in the kMC approach with the predicted microstructures. The chord length distributions method used in their workflow addresses the quantification of the grain size and shape distributions and their anisotropy in a microstructure. However, other important microstructural features, such as the volume fraction of the phase of interest, fail to be extracted by this method[18–21]. Fortunately, Fortunately, a rigorous quantitative framework based on the n-point statistics method has been developed to capture the statistical information of microstructure[22–25]. As the basis of the n−point statistics, the one-point statistics can reflect the probability density (i.e., volume fraction) of finding a specific discrete local state of interest at any randomly selected single point (or voxel) in a microstructure. Two-point statistics, a higher-order measurement, can capture the probability associated with finding an ordered pair of specific local states at the head and tail of a vector $\boldsymbol{t}$ that is randomly thrown into a microstructure. These statistics have been proven to contain unbiased and completed structural information[26,27]. However, an experimentally obtained microstructure with a small size always includes limited structure information and cannot be used as a representative volume element (RVE); it thus can hardly be associated with the macroscopic mechanical properties. One may emphasize a compromised scheme of com-

bining chemical compositions and experimental conditions to fill the gaps[7,9]. Unfortunately, this indirect strategy cannot essentially eliminate the statistical error caused by the small size of microstructures. Another surrogate solution is to approximate the statistics of an RVE by averaging that of multiple subdomains of the entire sample based on the assumption of statistical homogeneity. It is conceivable that more details of structure will be captured by these subdomains [a single domain refers to a statistical volume element (SVE) in this study] with a higher resolution. Niezgoda *et al.* proposed a novel concept called the RVE set consisting of a certain number of SVEs with the minimum size[28]. Through accessing the convergence of a quantitative metric D_s (root mean square error between individual statistics of SVE and the two-point statistics of a priori RVE or the average statistics of the overall SVEs), they successfully constructed such an optimal RVE set so that the distribution and dispersion of structural features match the entire material sample. This scheme has the advantages of saving time and computing resources for predicting mechanical properties by finite element analysis. Nevertheless, it is not applicable to establish PSP linkages in a real experimental situation without a prior RVE. The existing reports thus intuitively averaged the statistics of several SVEs to extract the maximum amount of structural information[29–32]. Therefore, it is necessary to explore new and universal methods for optimizing microstructure information to construct an RVE set and thus build a more reliable PSP linkage.

The effective information of two-point statistics is compressed in the central area after centralized transformation, leading to the statistics of an RVE containing a great deal of redundancy in the area with a large length of $\boldsymbol{t}$[7,33,34]. Determining the boundary of these two areas is beneficial for analyzing and understanding structure features, especially for the features related to length scales such as average grain size. Through an example of Al-alloy matrix composites, Tewari *et al.* found that numerous length parameters that characterize spatial heterogeneity and clustering of SiC particles can be extracted from two-point statistics[35]. Niezgoda *et al.* further defined a concept named coherence length, t_c, which is mathematically expressed as

$$\left\langle {}^{hh'}f_{\boldsymbol{t}}^{j} - {}^{h}f^{j} \cdot {}^{h'}f^{j} \right\rangle \leq \epsilon \quad \forall \|\boldsymbol{t}\| \geq t_c \tag{1}$$

where ${}^{hh'}f_{\boldsymbol{t}}^{j}$ represents the two-point cross-correlation statistics for the two local state h and h' of the jth members in an RVE set[22]. ${}^{h}f^{j}$ and ${}^{h'}f^{j}$ are their one-point statistics, namely volume fraction. $\langle\cdot\rangle$ denotes the ensemble average operation. The statistics in the area with $\boldsymbol{t}$ of length longer than t_c are considered as redundancy information. It can be imagined that the value of t_c will obviously change if ϵ is of a different magnitude, leading to an inaccurate measurement of the length scale associated with the structural features of interest. High dimensional redundant data may introduce unnecessary noise and impede the modeling of PSP linkage. However, there is no accurate reference value for this tolerance factor ϵ, and the existing studies are based on intuition to truncate redundancy[14,30,34,36,37]. Thus, determining the threshold of ϵ is also one of the important issues in optimizing microstructural information.

Principal component analysis (PCA)[38], a popular dimensionality reduction algorithm, can effectively address the above challenge of high-dimensional data and has been widely applied to many fields, such as grain coarsening[39], microstructure evolution during creep[31], nonmetallic inclusions in steels[37], *etc.* Interestingly, one can project statistical features of microstructures into a PCA space and compare them using some common distance metrics such as Euler distance[40–43], which provides a potential solution for optimizing microstructural information by constructing an RVE set and removing redundancy. More importantly, low-dimensional features of microstructures obtained by PCA can be input into machine learning (ML) models to establish high-fidelity PSP linkages[13,44–48].

In the present study, we developed a new scheme to build a more reliable structure-property (SP) linkage by optimizing microstructural information, which can be extended to a higher-ordered PSP linkage in the future. Taking an example of ferrite heat-resistant steels, we performed a series of experiments and built a small dataset. This kind of steel has become one of the main materials for the heavy and thick components of advanced ultra-supercritical (A-USC) power plants due to its high thermal diffusivity and low cost[49]. Significant efforts have

Table 1. Nominal chemical compositions and yield strength (σ) at 650 °C of the steels.The compositions of the elements Cr, C, Si, and Cu are identical for the five alloys and thus not listed (Cr, 15.00; C, 0.05; Si, 0.50; Cu, 0.10; unit, wt%)

Label	Mn	Ni	Al	Ti	Mo	W	σ(MPa)
Alloy 1	1.06	2.64	0.80	0.24	0.08	0.04	254
Alloy 2	0.96	3.20	1.12	0.08	0.24	0.08	299
Alloy 3	1.54	3.04	1.20	0.08	0	0.32	276
Alloy 4	0.60	3.12	1.12	0.08	0.32	0	336
Alloy 5	1.44	2.56	1.20	0	0.04	0.52	325

been made to understand the SP linkage and improve mechanical properties at a high temperature (650 °C) for the steels[50–53]. Using PCA and two-point statistics, we propose a new method and metric to construct an RVE set from the small SVEs of the steels. We also explored the effect of different redundancy-truncation levels of two-point statistics on the established ML model and determined the acceptable threshold of the tolerance factor ϵ. The reliability and generalization ability of this scheme were also proved by two other datasets including experimental data of Ni-Fe-based superalloys collected by Zhong *et al.*[54] and simulated data by phase field method (PFM), respectively.

MATERIALS AND METHODS

Materials preparation

Five alloys were prepared using the raw metals with purity higher than 99.99% by smelting, followed by casting into ingots of ≈ 40 g. The chemical compositions of the alloys are listed in Table 1. The samples were homogenized for 16 h at 1100 °C with subsequent air-cooling. Hot rolling was then performed at 1100 °C five times, each time holding for 10 min (60% final deformation). Heat treatment was achieved by austenitizing at 1100 °C for 0.5 h with posterior air cooling. The samples were then aged at 750 °C for 12 h, followed by air cooling. The microstructures of these alloys were characterized by optical microscopy (OM, Olympus P4000). High-temperature tensile tests at 650 °C were performed on a TSMT EM6.504 universal testing machine with a strain rate of $10^{-3}s^{-1}$. It is noted that the heat treatment was performed at 750 °C, and no phase transition occurred at 650 °C, so the microstructures could remain stable at 650 °C for a long time. Thus, the microstructures at room temperature were used to establish linkage with the yield strength at 650 °C.

Data preprocessing

All microstructures were binarized by an image processing technique named Otsu's threshold processing[55]. This technique is a nonparametric and unsupervised method of automatic threshold selection for picture segmentation. It selects a threshold automatically from a gray level histogram, and the threshold is equal to the one specific pixel value $f(i)$, which is determined by the maximum variance of the foreground and background pixels[56]. Moreover, the discrete microstructures were transformed to a uniform dimension using the transform.rescale(·) function in the scikit-image library to ensure that one pixel corresponds to an actual size of $0.6504\mu m$, as shown in Figure 1. It can be seen that the austenite phase and ferrite phase in the original microstructure are well separated by black and white pixels, and the noise points (gray texture) are completely eliminated.

To eliminate the impact of data magnitude differences on the performance of the model, before establishing the SP linkage by a ML model, all of the microstructural features (low-dimensional representativeness of the microstructures) $\boldsymbol{x}$ were normalized using $\widetilde{\boldsymbol{x}} = (\boldsymbol{x} - \mu)/\varepsilon$, where μ and ε denote the mean and variance, respectively.

Extracting microstructural information

The circular sample, as displayed in Figure 2A, may rotate during OM characterization, leading to inconsistent statistics of the same microstructure in different reference frames. To filter out the dependence of the statistics

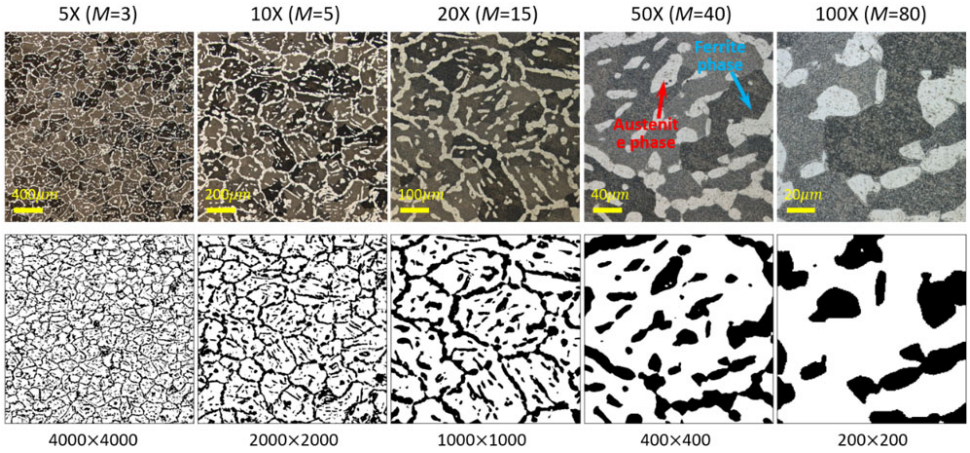

Figure 1. Experimental microstructures and corresponding binary results for Alloy 1. The magnification of the microscope and the number of microstructures at each magnification is listed at the top, while the grid number is given at the bottom.

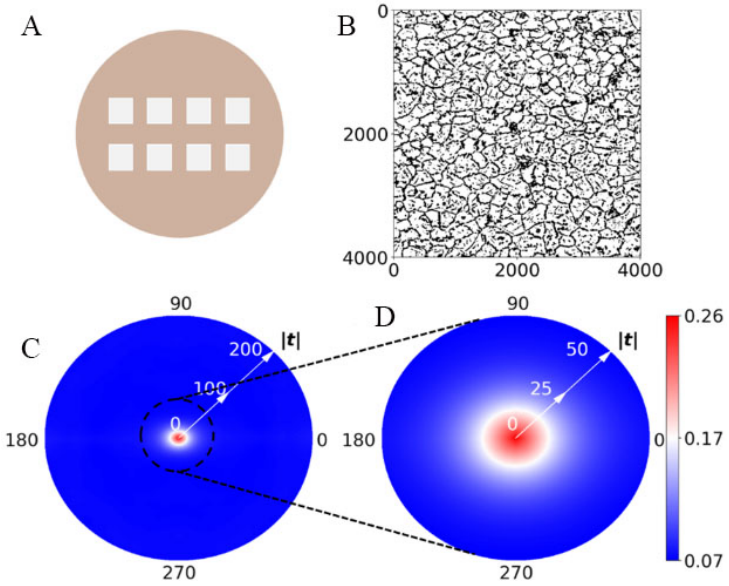

Figure 2. Example of a microstructure and corresponding autocorrelations: (A) schematic diagram of sample preparation; (B) binarized micrograph; (C) autocorrelations of the black phase; and (D) enlarged drawing of the central domain in (C) labeled by the dotted black circle.

on the observer reference frame, we employed rotationally invariant two-point statistics (RI2SS) to capture the important structural details[23]. For a certain local state h (ferrite phase, $h = 0$; austenite phase, $h = 1$), the microstructure function m_s^h representing the volume fraction of local state h in the location of $\boldsymbol{s}$ should be calculated firstly. The two-point statistics is mathematically expressed as

$$f_t^{hh'} = \frac{1}{|S_t|} \sum_s m_s^h m_{s+t}^{h'} \tag{2}$$

where $\boldsymbol{t}$ is the discretized vector placed in microstructure and $|\boldsymbol{S}_t|$ denotes the total number of valid trials associated with discrete vector $\boldsymbol{t}$. In this work, we only calculated two-point autocorrelation statistics of the black phase in the microstructure, as shown in Figure 2B, which can be obtained when $h = h'$. Notedly, the RI2SS of the microstructure is further calculated, as shown in Figure 2C and D. For convenience, all of the statistics are referred to as autocorrelations. The peak value of Figure 2C represents the volume fraction of

the target phase, and the main spatial features (average size and shape distribution of the phase, *etc.*) of the microstructure are contained in the central area that is enlarged in Figure 2D. In addition, the invariant value in the blue area is approximately equal to the square of the peak and represents redundant information.

PCA was used to reduce the dimensionality of autocorrelations. One can obtain principal component scores (PCs), i.e., low-dimensional features of a microstructure, through projecting its autocorrelations into a new space supported by several orthogonal basis vectors. The vectors are ordered and selected according to their explained variance that reflects the main variation of the samples. Mathematically, the original autocorrelations can be reconstructed by

$$f_t^{11,(j)} = \sum_{i=1}^{min\{(J-1),R\}} \alpha_i^{(j)} \phi_i^{11} + \bar{f}_t \tag{3}$$

where $f_t^{11,(j)}$ represents the autocorrelations of the jth microstructure. $\bar{f}_t = \frac{1}{J}\sum_{j=1}^{J} f_t^{11,(j)}$, where $\bar{f}_t$ and J denote the ensemble average and number of all of the autocorrelations. $\alpha_i^{(j)}$ and ϕ_i^{11} represent the ith PCs of the jth member and the ith basis vector, respectively. R is the dimensionality of autocorrelations. As the main parameter, α_i participates in the subsequent analysis and modeling.

Modeling and evaluation

We used a classical ML regression model, Ridge regression [57], to build the SP linkage. By imposing the penalty $\alpha\|\boldsymbol{\omega}\|_2^2$, Ridge can solve some problems of ordinary least squares. Mathematically, the objective function of Ridge is to minimize a penalized residual sum of squares:

$$\min_{\boldsymbol{\omega}} \|\boldsymbol{x}\boldsymbol{\omega} - \boldsymbol{y}\|_2^2 + \alpha\|\boldsymbol{\omega}\|_2^2 \tag{4}$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are the inputted features and outputted yield strength in this study. α is the complexity parameter that controls the amount of shrinkage: the larger is the value of α, the greater is the amount of shrinkage. Thus, the coefficients become more robust to collinearity. $\boldsymbol{\omega}$ represents the coefficients of $\boldsymbol{x}$. The Ridge models in this work were trained by calling the scikit-learn library in Python 3.7 [58]. All of the models keep the default hyperparameters.

The performance of these models was quantified by root mean square error ($RMSE$) and determined coefficient (R^2), which are given as $RMSE = \sqrt{\frac{1}{M}\sum_{i=1}^{m}(y_i - \widehat{y_i})^2}$ and $R^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2}$, where y_i and $\widehat{y_i}$ are the experimental and predicted yield strength, respectively, and $\bar{y} = \frac{1}{M}\sum_{i=1}^{m} y_i$ denotes the average of M samples. The smaller the $RMSE$ is, and the closer R^2 gets to 1, the higher the prediction accuracy. In addition, the leave-one-out cross-validation (LOOCV) approach was employed to evaluate $RMSE$ and R^2.

SCHEME

We propose a data-driven scheme for building SP linkage including five modularity: dataset preparation, data preprocessing, microstructural information extraction, microstructural information optimization, and SP linkage construction [Figure 3]. All parameters and corresponding explanations are listed in Table 2. The details of applying this scheme to the Ferrite steels are as follows:

1) Creating experimental dataset $\boldsymbol{D}(\boldsymbol{x}_{\boldsymbol{S},\boldsymbol{M}}, \boldsymbol{y})$ by following the procedure mentioned above. The subscripts $\boldsymbol{S}$ and $\boldsymbol{M}$ label the size of the microstructures and the number of SVEs under different S. Here, $\boldsymbol{S} = \{S|S_{5X}, S_{10X}, S_{20X}, S_{50X}, S_{100X}\}$, $\boldsymbol{M} = \{M_S|3, 5, 15, 40, 80\}$. The dataset includes 5×143 microstructures and yield strength for the five alloys. For distinguishment, M_S denotes the maximum number of samples in a subset with a specific size S, and m represents the number of randomly selected samples in the subset, here $m \leq M_S$.

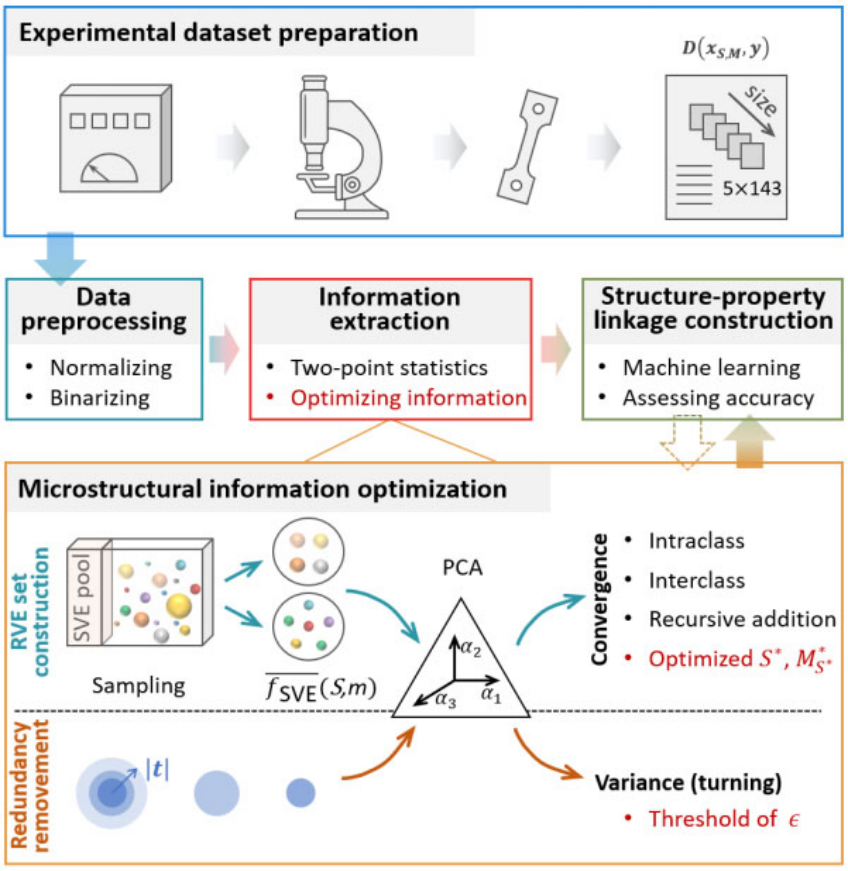

Figure 3. Scheme of optimizing microstructural information and constructing SP linkage. Note that the arrows with a colored filling point to the modeling path of SP linkage, and the arrow with dotted lines represents that the ML modeling is also used to determine the optimized S^*, $M_{S^*}^*$, and the threshold of ϵ. SP: Construction of the structure-property; PCA: principal component analysis.

2) Preprocessing microstructure and property data by Otsu's threshold processing and normalization operation mentioned above.

3) Extracting quantitative information of all microstructures by RI2SS.

4) Optimizing microstructural information to represent the structural features in the whole sample for each alloy. This procedure includes two sub-paths labeled by the colored arrows in the orange box in Figure 3:

a) Constructing RVE set (confirming the size and number of the included SVEs). We randomly selected different numbers m of SVEs with a certain size S to form a subset, calculated their average autocorrelations (simple arithmetic average of the autocorrelations of the all SVEs) $\overline{f_{SVE}}(S,m)$, and then projected all possible $\overline{f_{SVE}}$ into a PCA space to obtain corresponding low-dimensional features $\overline{\alpha_S^m}$. The two distances ($\overline{d_{S,centroid}^{nor}}$ and $\overline{d_{S,target}^{nor}}$) expressed by Equation (5) and (6) were next calculated, and the convergence along different S (interclass) and m (intraclass) was assessed to confirm the optimal size S^* of SVEs for constructing an RVE set. It is easy to understand that the locations of the SVEs with S larger than S^* will be clustered in the PCA space.

$$\overline{d_{S,centroid}^{nor}} = \frac{1}{m}\sum_{j=1}^{m}\sqrt{\sum_{i=1}^{3}\frac{\left(\overline{\alpha_{i,S}^m} - \frac{1}{m}\sum_{m}^{M_S}\overline{\alpha_S^m}\right)^2}{\left(\alpha_{i,max}-\alpha_{i,min}\right)^2}} \tag{5}$$

Table 2. Parameters used in this study and corresponding explanation list

Parameters	Explanation
$\boldsymbol{t}$	Random vector thrown into a microstructure
h, h'	Local state of interest (austenite phase and ferrite phase in our case)
${}^{h}f^{j}$	One-point statistics of local state h
$\boldsymbol{s}$	Cell node indexed the spatial domain of a microstructure
$\lvert \boldsymbol{S_t} \rvert$	The total number of valid trials associated with discrete vector $\boldsymbol{t}$
$m_{\boldsymbol{s}}^{h}$	Microstructure function representing the volume fraction of local state h in the location of $\boldsymbol{s}$
$f_{\boldsymbol{t}}^{hh'}$	Two-point cross-correlation statistics of local state h and h'
$f_{\boldsymbol{t}}^{hh}$	Two-point autocorrelation statistics of local state h and h'
t_c	Coherence length of a realistic structure
ϵ	Tolerance factor related to redundant information in two-point statistics.
j	The number of a microstructure sample or two-point statistics in a set
$\boldsymbol{J}$	The number of all two-point statistics in a set
$f(i)$	Pixel value in a digital microstructure
$\boldsymbol{x}$	Inputted feature array of the ML model
$\boldsymbol{y}$	Outputted property array of the ML model
$\hat{y}_i$	Predicted property of the ith sample by the ML model
$\bar{y}$	The average property of the samples
μ	Mean of inputted features
ε	Variance of inputted features
σ	Yield strength at 650°C
ω	The coefficients of $\boldsymbol{x}$ fitted by Ridge model
α	The complexity parameter that controls the amount of shrinkage in Ridge model
$\alpha_i^{(j)}$	The ith PC score of the jth SVEs
ϕ_i^{11}	The ith PC basis vector
$\bar{f}_{\boldsymbol{t}}$	The ensemble average of the two-point statistics
$\boldsymbol{R}$	The dimensionality of the two-point statistics
$\boldsymbol{D}$	The experimental dataset in this study
$\boldsymbol{S}$	The set of true size of the SVEs, i.e., $\boldsymbol{S} = \{S \lvert S_{5X}, S_{10X}, S_{20X}, S_{50X}, S_{100X}\}$
S^*	The optimal size of the SVEs in an RVE set
$\boldsymbol{M}$	The set of the number of SVEs with different sizes, i.e., $\boldsymbol{M} = \{M_S \lvert 3, 5, 15, 40, 80\}$
$M_{S^*}^*$	The optimal number of the SVEs in an RVE set
m	The number of randomly selected samples in the subset, here $m \leq M_S$
n	The number of primary grains in the PFM model
γ_4	Anisotropy coefficient of the solid-liquid interfacial energy in PFM model
ΔT	Nucleation supercooling in PFM model
$P_{\boldsymbol{t}}$	Pair correlation function of a microstructure
D_s	Root-mean-square error between the two-point statistics of each SVE and the target ensemble-averaged statistics

$$\overline{d_{S,target}^{nor}} = \frac{1}{m}\sum_{j=1}^{m}\sqrt{\sum_{i=1}^{3}\frac{\left(\overline{\alpha_{i,S}^{m}} - \overline{\alpha_{S_{5X}}^{3}}\right)^2}{\left(\alpha_{i,max} - \alpha_{i,min}\right)^2}} \tag{6}$$

where $\alpha_{i,max} = \max\{\alpha_{i,S}^{(1)}, \cdots, \alpha_{i,S}^{(m)}\}$, $\alpha_{i,min} = \min\{\alpha_{i,S}^{(1)}, \cdots, \alpha_{i,S}^{(m)}\}$. The smaller $\overline{d_{S,centroid}^{nor}}$ is, the closer the position of the autocorrelations of the SVEs in PCA space is to that of their ensemble average. A similar relationship applies to $\overline{d_{S,target}^{nor}}$, except that the object of comparison becomes the target autocorrelations that are obtained by averaging autocorrelations of the large domains with a size of S_{5X}. It is noted that these large

domains here were selected because they contained sufficient structural features that are independent of their location in the sample, as shown in Figure 1.

We also propose a novel method called "recursive addition" to confirm the optimal number $M_{S^*}^*$ of SVEs for constructing an RVE set. When one gradually introduces a new SVE, the average autocorrelations will also include more and more structural information. If the diversity of the structural features in these SVEs is saturated enough to match the entire material sample, the locations of the $\overline{f_{SVE}}(S^*, m)$ will gather in a small area in the PCA space. In other words, the distance between two adjacent points in the space will stabilize around a sufficiently small value. This distance can be mathematically expressed as

$$d_{\left|\overline{\alpha_{S^*}^J} - \overline{\alpha_{S^*}^{J-1}}\right|} = \sqrt{\sum_{i=1}^{3} \frac{\left(\overline{\alpha_{i,S^*}^J} - \overline{\alpha_{i,S^*}^{J-1}}\right)^2}{\left(\alpha_{i,max} - \alpha_{i,min}\right)^2}} \tag{7}$$

b) Removing redundant information of the autocorrelations. We truncated the autocorrelations in the constructed RVE set by controlling the different maximum lengths of $\boldsymbol{t}$ and then observed the variation of their variance in PCA space to explore the threshold of ϵ defined by Equation (1). Finally, the structural information that contains the least redundancy is retained.

5) Establishing SP linkage. By inputting the low-dimensional features of the RVEs for the five alloys, we trained a Ridge regression model and assessed its accuracy in predicting yield strength. It is noted that this process was also used to validate the reliability of the methods proposed in Procedure (4).

This scheme shown in Figure 3 was also performed on the dendrite solidification data from PFM for Al-Cu alloys and experimental data of Ni-Fe-based superalloys. The reliability and generalization ability of the scheme were also considered in this study.

RESULTS

Construction of RVE set

Following the workflow shown in Figure 3, we traversed all possible combinations with the variation of S and m in the SVE pool, calculated their average autocorrelations $\overline{f_{SVE}}(S, m)$, and then employed PCA to extract their low-dimensional features $\overline{\alpha_S^m}$, where $S \in \boldsymbol{S}, m \leq M_S$, and $M_S \in \boldsymbol{M}$. It is noted that all of the features were grouped into five clusters according to the different sizes of SVEs, and their distributions in the PCA space are shown in Figure 4. It can be observed that, for a small S (S_{50X} or S_{100X}), the larger m is, the more concentrated the distribution of sample points are and the further away $\overline{\alpha_S^m}$ is from $\overline{\alpha_{S_{5X}}^3}$, indicating that more structural features are included, but still not enough to match that of a larger microstructure. For a large S (S_{5X}, S_{10X}, or S_{20X}), all of the points appear to be centrally distributed in a small region, which demonstrates that the structural diversity included in the extracted information represents a saturation.

To quantify the interclass and intraclass convergence observed above, we used Equation (5) and (6) to calculate the normalized average distances, $\overline{d_{centroid}^{nor}}$ and $\overline{d_{target}^{nor}}$. Figure 5A explains the calculation principle of the single distance $\overline{d_J^{nor}}$; the red star point named target is associated with $\overline{\alpha_{S_{5X}}^3}$. The variation of these two distances with the size and number of SVEs is given in Figure 5B. It can be seen that $\overline{d_{target}^{nor}}$ quickly declines and gradually converges in the range of less than 0.1 as the size of SVEs increases. Thus, the minimum size of SVE that can be used to construct an RVE set is determined as $S^* = S_{20X}$. When $S > S_{20X}$, the decrease rate of $\overline{d_{centroid}^{nor}}$ is first fast, then slow, and finally approaches 0, indicating that the structural information contained in the SVEs reaches saturation. However, what needs to be emphasized is that the calculation for $\overline{d_{centroid}^{nor}}$ is based on a premise of the SVE pool, which is inconsistent with the requirement of low cost and the fact that there are only several SVEs in experiments. Therefore, the volume of the RVE set needs to be determined separately.

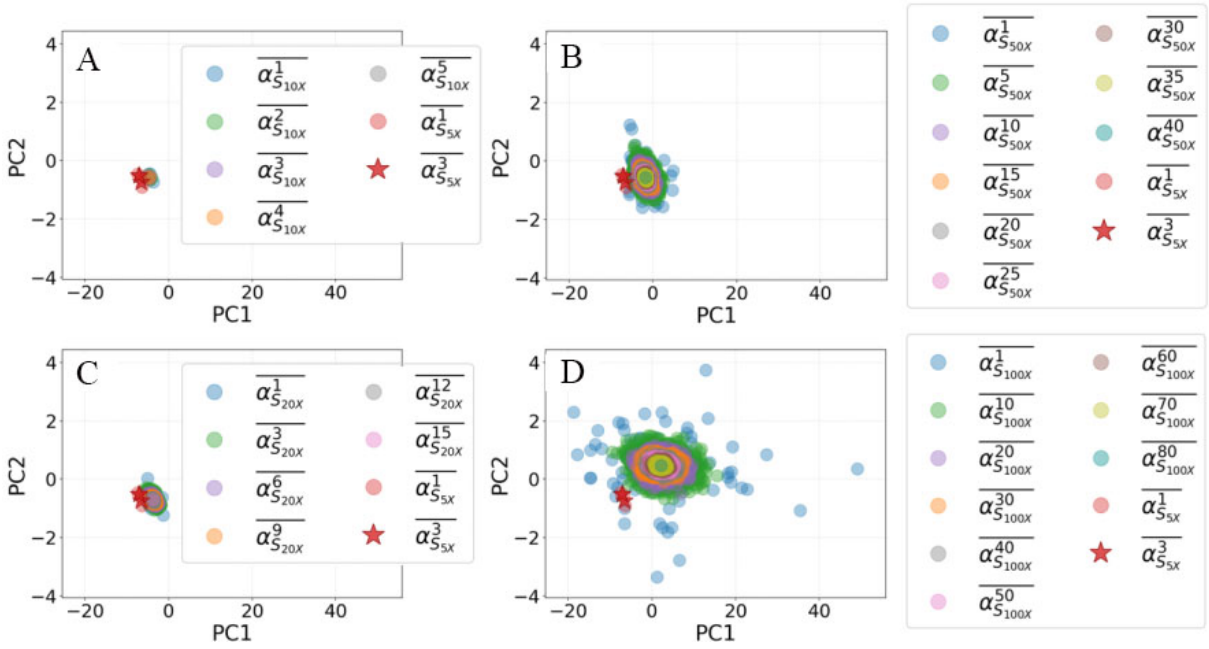

Figure 4. The distribution of the averaged autocorrelations of a certain number SVEs in the PCA space. The PCs are grouped into five clusters that are distinguished by the size S of these SVEs: (A) S_{5X} and S_{10X}; (B) S_{20X}; (C) S_{50X}; and (D) S_{100X}.

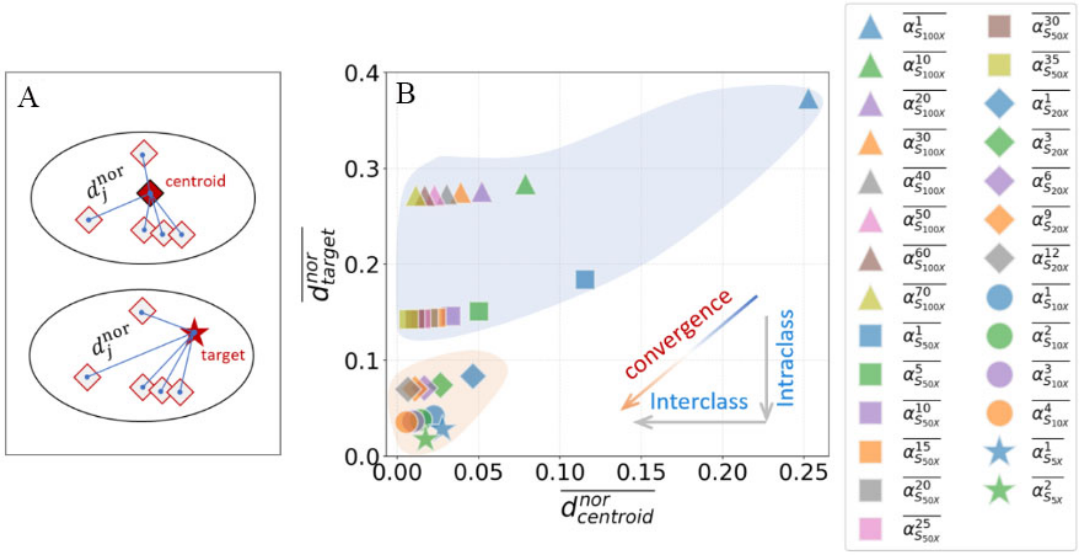

Figure 5. Evaluation of convergence and reliability of ensemble averaged autocorrelations: (A) schematic diagram of evaluation metrics, $\overline{d^{nor}_{centroid}}$ and $\overline{d^{nor}_{J,target}}$; and (B) comparison between the two averaged metrics.

As for the volume of the RVE set, we propose a novel method named recursive addition based on Euler distance in the PCA space. Figure 6B explains the rationality of the method by using a defined distance, $d_{\left|\overline{\alpha^{J}_{S^*}}-\overline{\alpha^{J-1}_{S^*}}\right|}$, which is mathematically expressed by Equation (7). Here, S^* represents the optimal size of the SVEs in the RVE set, while $\boldsymbol{\alpha}^{j}_{S^*}$ and $\boldsymbol{\alpha}^{j-1}_{S^*}$ are the PC features of the jth and $(j-1)$th SVEs added gradually. It is easy to understand that $d_{\left|\overline{\alpha^{J}_{S^*}}-\overline{\alpha^{J-1}_{S^*}}\right|}$ will gradually decrease and eventually stabilize in an acceptable range when the new members are added continuously, as shown in the green shaded area in Figure 6B. We started from the five SVEs with the size of S_{20X} for the alloys, then ran PCA on the autocorrelations of these microstructures to

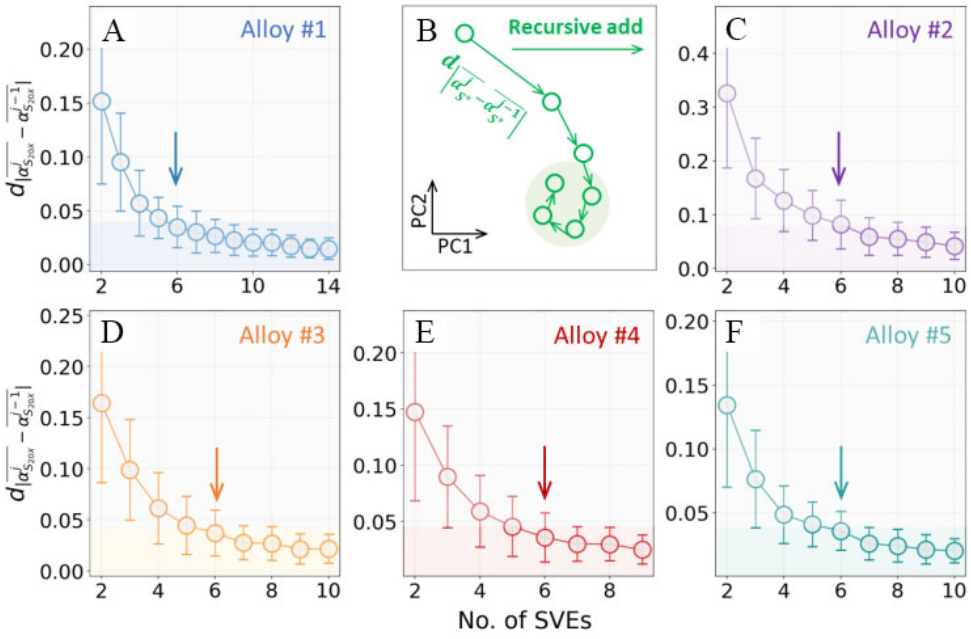

Figure 6. (A,C-F) Distance $d_{\left|\overline{\alpha^{J}_{S^*}}-\overline{\alpha^{J-1}_{S^*}}\right|}$ as a function of the number of SVEs by recursively adding five experimental samples for Alloys 1-5; (B) the application diagram of the distance.

obtain low-dimensional features $\alpha^{0}_{S_{20X}}$, and subsequently projected a new SVE into the PCA space to get $\alpha^{1}_{S_{20X}}$. We next calculated $d_{\left|\overline{\alpha^{1}_{S_{20X}}}-\overline{\alpha^{0}_{S_{20X}}}\right|}$, repeating the above steps several times for each alloy. To produce reliable and non-random results, we performed this recursive addition method 100 times to assess the mean and standard deviation of $d_{\left|\overline{\alpha^{1}_{S_{20X}}}-\overline{\alpha^{0}_{S_{20X}}}\right|}$. Figure 6(A, C-F) displays the variation of the distance as the number of additions increases for the five alloys. It can be found that $d_{\left|\overline{\alpha^{J}_{S_{20X}}}-\overline{\alpha^{J-1}_{S_{20X}}}\right|}$ quickly declines followed by a slow reduction. The distances for the alloys finally converge to less than 0.05. As pointed from the vertical arrows shown in the Figures, we determined to use six members to construct an RVE set, and the structural features contained in the set can consistently represent that of the whole sample.

Construction of structure-property linkage

We then employed Ridge regression to extract SP linkage. The inputs of the model are the low-dimensional features ($\overline{\alpha^{6}_{S_{20X}}}$) of the constructed RVE set for the five experimental alloys, and the output is yield strength. LOOCV was used to assess the prediction accuracy. For comparison, we also built three other Ridge models using different sets of inputs ($\overline{\alpha^{2}_{S_{20X}}}$, $\overline{\alpha^{10}_{S_{20X}}}$, and $\overline{\alpha^{3}_{S_{5X}}}$) obtained from the average autocorrelations of 2 and 10 SVEs with size of S_{20X} and 3 SVEs with size of S_{5X}, respectively. To avoid randomness, the selection procedure of these SVEs was repeated 100 times. Figure 7A and B exhibits the distribution of $\overline{\alpha^{2}_{S_{20X}}}$, $\overline{\alpha^{6}_{S_{20X}}}$, $\overline{\alpha^{10}_{S_{20X}}}$ (hollow points), and $\overline{\alpha^{3}_{S_{5X}}}$ (solid points) in the PCA space. It can be observed that there is always a hollow point occupying a position further away from the solid point for each alloy. After verification, we found that this isolated point is associated with $\overline{\alpha^{2}_{S_{20X}}}$, which is consistent with the results shown in Figure 6. In Figure 7C, it can be seen that the accuracies of the models with the input of $\overline{\alpha^{6}_{S_{20X}}}$, $\overline{\alpha^{10}_{S_{20X}}}$, and $\overline{\alpha^{3}_{S_{5X}}}$ are extremely close to each other (R^2 are 0.8430, 0.8680, and 0.8652, respectively), which is improved by at least 28.57% compared with the model inputting $\alpha^{2}_{S_{20X}}$. Figure 7D compares the experimentally measured yield strength and the ones predicted by the model inputting $\overline{\alpha^{6}_{S_{20X}}}$ from the RVE set. The diagonal distribution between them also indicates the high accuracy of the model. The mean absolute error (MAE) is less than 10 MPa (embedding subgraph in Figure 7D), which demonstrates that the structural information contained in our constructed RVE

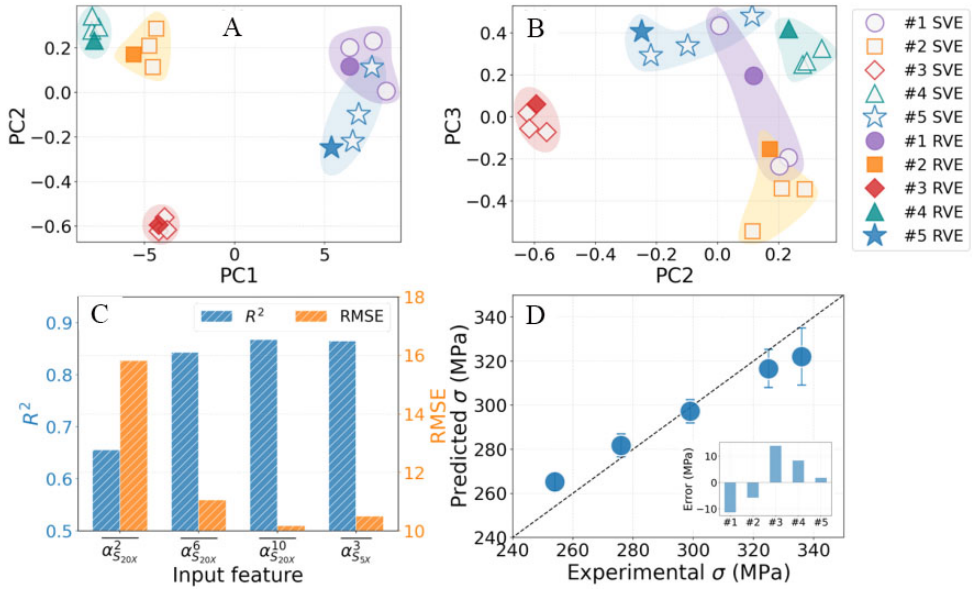

Figure 7. Distribution of the alloys #1-#5 in the PCA space and prediction accuracy (R^2 and $RMSE$) for the yield strength: (A) distribution along PC1 and PC2 vectors; (B) distribution along PC2 and PC3 vectors; (C) R^2 and $RMSE$; and (D) comparison between predictions and experiments. The embedding sub-plot in (D) shows their difference for the five alloys.

set can be mapped to the macroscopic mechanical property of the whole sample.

To verify the generalization ability of the proposed method in constructing the RVE set, we performed this method on a dataset of dendrite solidification of Al-Cu alloys simulated by PFM [59–63]. The parameters of PFM are listed in Table S1. The dataset includes 48 microstructures that are produced by controlling solidification parameters including the number of primary grains (n), anisotropy coefficient of the solid-liquid interfacial energy (γ_4), and nucleation supercooling (ΔT). From the results shown in Figure S1, it can be observed that the difference of these microstructures comes from the grain morphology and volume fraction of the solid phase. RI2SS and PCA were then employed to extract their average autocorrelations and low-dimensional features. The distribution of the features is shown in Figure S2. Combined with the results of Figures S1 and S2, we demonstrated that the microstructures distinguished by ΔT and γ_4 placed along PC1 and PC2 vectors, respectively, indicating that the first two PCs reflect the variation of volume fraction and grain morphology. Through applying the recursive addition method, as shown in Figure S3, an RVE set consisting of six SVEs was constructed. The established PS linkage shown in Figure S4 also reveals the reliability of this RVE set in extracting sufficient structural features. More importantly, the successful application of the proposed method on the simulation dataset proves its credibility and universality.

Identification of Redundant statistics

Two-point statistical autocorrelations contain valuable information concentrated in the central area and vast redundant information in the peripheral area. Following the procedures shown in Figure 3, we truncated the autocorrelations of the microstructures in the RVE set, as displayed in Figure 8A. The maximum modulus of the vector $\boldsymbol{t}$ is labeled as $|\boldsymbol{t}|\,max$. Using the truncated autocorrelations with a certain $|\boldsymbol{t}|_{max}$ (20-100 pixels) for the five alloys, we created a PCA space and projected these statistics into the space, and then examined the variation of PC variance, as shown in Figure 8B and C. As $|\boldsymbol{t}|_{max}$ decreases, the cumulative variance of the first three PCs does not change significantly and that of the first two PCs increases slightly. When $|\boldsymbol{t}|_{max}$ reduces from 100 to 50 pixels, the individual variance of PC1 declines slowly and that of PC2 rapidly rises. When $|\boldsymbol{t}|_{max}$ continues to be reduced, their tendencies reverse. Different from the first two PCs, the trend of individual variance of PC3 is inconsistent with that of $|\boldsymbol{t}|_{max}$. Combining with Equation (3), we further investigated

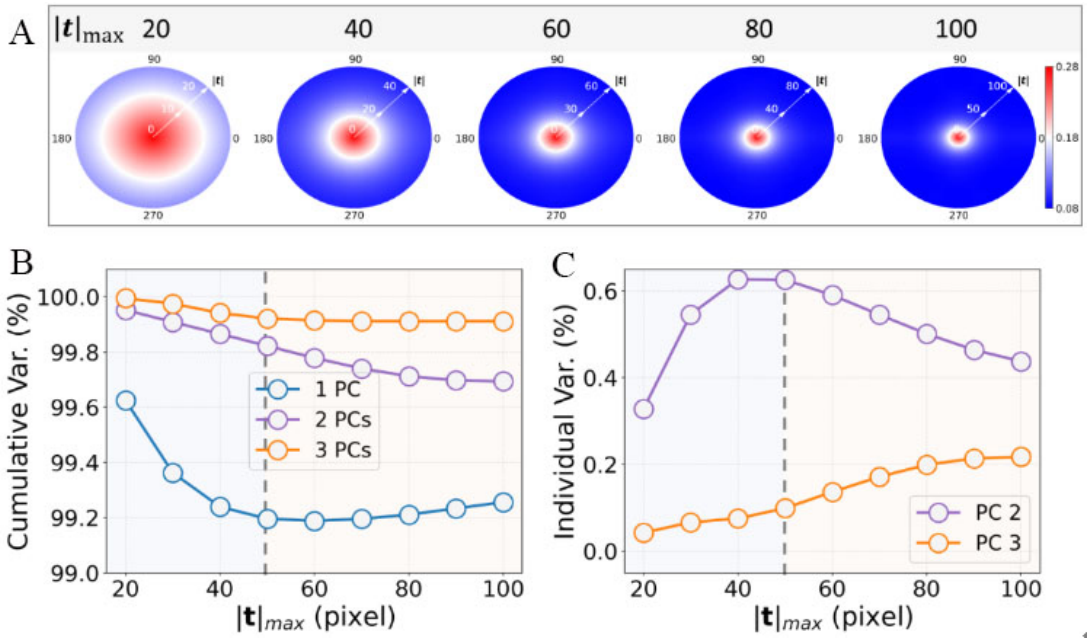

Figure 8. Variation of autocorrelations and PC variance with different lengths of $|t|_{max}$: (A) the autocorrelations of Alloy 1; (B) cumulative variance for the first three PC3; and (C) individual variance for PC2 and PC3.

variation of the first PC basis vectors (ϕ_1^{11} and ϕ_2^{11}) with the truncation of $|t|_{max}$. Figure S5 demonstrates that, when $|t|_{max} \geq 50$ pixels, if PC1 increases, the peak value of the autocorrelations that is strongly associated with the volume fraction of the austenite phase will also increase. In other words, PC1 reflects the volume fraction. As for PC2, it mainly relates to the peak value and the size of the central area, indicating that PC2 determines the volume fraction, average size, and distribution of the austenite phase. When $|t|_{max} < 50$ pixels, PC1 is not only correlated with phase volume fraction but also related to the average size of the phase, and PC2 does not contain the information about phase distribution as it does before. Therefore, we hypothesized that the effective information in the autocorrelations is removed when $|t|_{max} < 50$ pixels, leading to a change in physical meaning of the low-dimensional features and a mutation in PC variance.

While removing redundancy in statistical autocorrelations, the distribution of the SVEs in the RVE set in the PCA space was also altered, as shown in Figure S6. When $|t|_{max} \geq 50$ pixels, the low-dimensional features hardly change during truncating while obvious changes of them can be observed in the case of $|t|_{max} < 50$ pixels. We quantified the distances between these low-dimensional points and their centroid for each alloy, and then the variance of the distances was labeled as intraclass variance, while the interclass variance represented that between the centroids for the five alloys. Figure 9A visualizes the two variances as a function of $|t|_{max}$. When $|t|_{max}$ reduces from 100 to 50 pixels, intraclass variance slightly decreases and interclass variance varies in an inverse tendency, as indicated by the dotted arrows. It is easily understood that the discrepancy of the curves shown in Figure 9B within the same class (a certain alloy) is dominated by the external redundancy included in the autocorrelations compared with that in different classes (different alloys), whose difference is mainly determined by the central areas of the autocorrelations. The two variances compete with each other, resulting in little variation in the overall population (orange line in Figure 8B). When $|t|_{max} < 50$ pixels, the valuable information in the central area starts to be eliminated, the interclass and intraclass variances both increase, and the overall variation is also intensified (orange line in Figure 8B). Therefore, these results prove our hypothesis above, i.e., the critical length of $|t|$ that distinguishes valuable information and redundant information is 50 pixels for our microstructures.

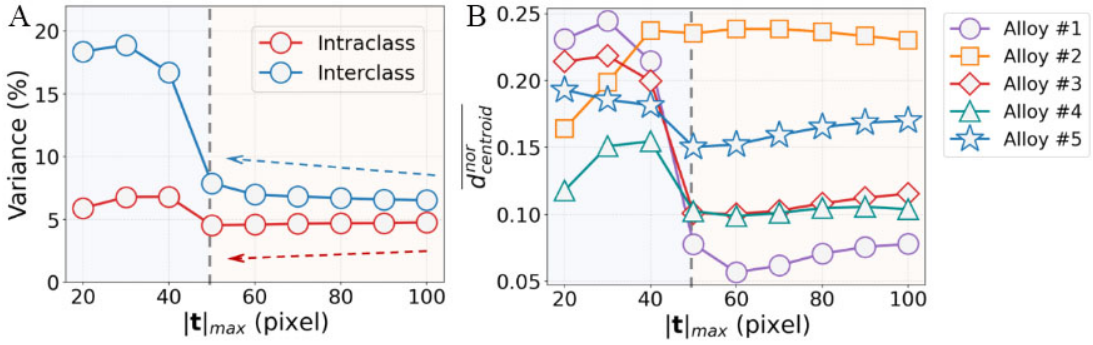

Figure 9. Variation of intraclass variance, interclass variance and $\overline{d_{centroid}^{nor}}$ with different lengths of $|t|_{max}$ for the five alloys: (A) intraclass variance and interclass variance; and (B) $\overline{d_{centroid}^{nor}}$. Three quantitative metrics were calculated from the scatter plots of the low-dimensional features, as shown in Figure S6.

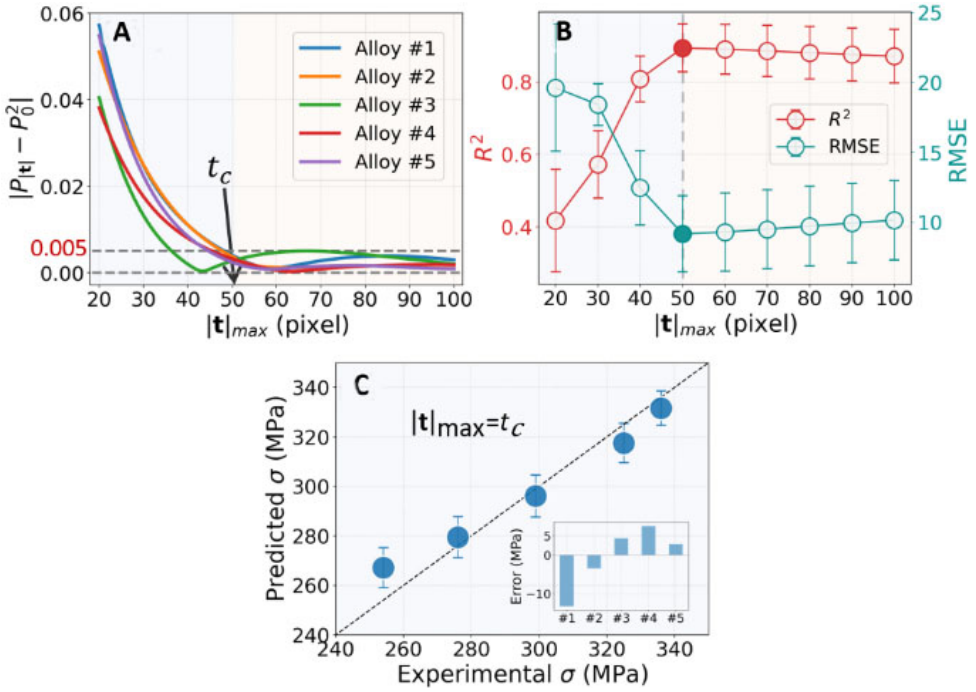

Figure 10. Variation of $|P_{|t|} - P_0^2|$ and the model accuracy with the change of the different lengths of $|t|_{max}$: (A) $|P_{|t|} - P_0^2|$; (B) R^2 and $RMSE$; and (C) comparison between predictions and experiments. The embedded subgraph shows that the prediction errors of the five alloys are within ±6.28 MPa.

Improvement of structure-property linkage

Tolerance factor ϵ defines a length scale feature of microstructure called coherence length t_c by Equation (1). However, the certain threshold of ϵ is still unknown. An excessively large ϵ will mislead the choice of t_c and may lead to a failure of SP linkage. This section is mainly devoted to confirming a precise threshold to improve the built SP linkage.

By calculating pair correlation function (PCF) of the average autocorrelations in the RVE set, we modified the left-hand side of Equation (1) as $\langle P_{|t|}^{(j)} - (P_0^{(j)})^2 \rangle$ [23]. For convenience, the item is simply expressed as $\left|P_{|t|} - P_0^2\right|$. Obviously, it is a function of $|t|$. Figure 10A gives the variation of $\left|P_{|t|} - P_0^2\right|$ as a function of $|t|_{max}$ for the five alloys. When $|t|_{max} \geq 50$ pixels, all of the curves present a plateau. At this point, the values of $\left|P_{|t|} - P_0^2\right|$ are less than a threshold of 0.005; thus, the critical t_c was confirmed to be 50 pixels (32.52 μm).

Using the autocorrelations with different $|\boldsymbol{t}|_{max}$ (20-100 pixels), we established several SP linkages by Ridge regression and employed LOOCV to evaluate their accuracies, as shown in Figure 10B. $RMSE$ of the model for $|\boldsymbol{t}|_{max} = t_c$ reduces by 9.67% compared with that for $|\boldsymbol{t}|_{max} = 100$ pixels (no truncation), and R^2 increases by 2.62%. When $|\boldsymbol{t}|_{max} < t_c$, the performance of the models starts to deteriorate. The results of the best model are highlighted by red and cyan solid points in Figure 10B, and the predictions agree well with the experiments shown in Figure 10C. The embedded subgraph shows that the MAE between the predicted value and the experimental one is within 6.28 MPa, which is reduced by 37.2% compared with the results in Figure 7D.

We further employed the procedure in Figure 3 on a Ni-Fe-based superalloy dataset to explore the impact of redundancy removement on the accuracy of SP linkage and the threshold ϵ. The dataset was collected from [54]. It is noted that the microstructures shown in Figure S7 are extremely different from our experimental ones shown in Figure 1 in terms of morphology. Extraction of low-dimensional features and construction of SP linkage on this superalloy dataset are visualized in Figures S8 and S9. Eventually, the accuracy of the models along with the change of $|\boldsymbol{t}|_{max}$ trends similar to Figure 10B, and the threshold ϵ that is used to determine t_c is also less than 0.005, demonstrating the reliability and university of this threshold in confirming the coherence length of experimental microstructure and assisting in establishing SP linkages.

DISCUSSION

Advantages of the quantitative metrics based on PCA

Quantitative comparison between two microstructures has always been a fascinating issue. To complete this task, Niezgoda *et al.* developed a metric D_s to reflect the root-mean-square error between the two-point statistics of each SVE and the target ensemble-averaged statistics, where s is the size of the selected SVEs [28]. When the errors for all SVEs are small and close to each other, the amount of information included in the two-point statistics of these SVEs will be saturated and independent of the size and number of the SVEs. Niezgoda *et al.* used D_s to successfully construct an RVE set that can be used to predict mechanical properties in a computationally economical manner [22]. Figure S10E provides the variation of D_s with the change of $|\boldsymbol{t}|_{max}$ in the autocorrelations. Obviously, D_s is strongly correlated with $|\boldsymbol{t}|_{max}$. In other words, even for the same group of microstructures, D_s fails to provide a specific and valuable reference for different operators. In addition, as for two SVEs $m^{(j)}$ and $m^{(k)}$ located at different spatial positions of the same sample, the two elements $f_{\boldsymbol{s}_0}^{(j)}$ and $f_{\boldsymbol{s}_0}^{(k)}$ in their autocorrelations $f_{\boldsymbol{t}}^{(j)}$ and $f_{\boldsymbol{t}}^{(k)}$ have no specific physical meanings except that the peak values $f_{\boldsymbol{0}}^{(j)}$ and $f_{\boldsymbol{0}}^{(k)}$ represent volume fraction of the local state of interest, where the superscripts j and k label the two SVEs. Therefore, the error metric produced by two autocorrelations can only reflect the average degree of similarity in morphology pattern of them, but it fails to rigorously measure the distinguishment in physical features of the microstructures.

Our proposed quantitative metrics based on PCA successfully overcome the defects above. From Equation (5) - (7), we can find that the metrics are distance measurement between the low-dimensional features $\alpha_i^{(j)}$ and $\alpha_i^{(k)}$ of $m^{(j)}$ and $m^{(k)}$. Generally, the first several PCs have interpretable physical meanings [29,34,43]. For the constructed RVE set in the present study, α_1 represents the volume fraction of the austenitic phase and α_2 quantifies the average size and distribution of this phase (a detailed understanding is provided in Figure S5). Thus, the metrics based on PCA can rigorously measure the degree of similarity in physical features of microstructures. In addition, when $|\boldsymbol{t}|_{max} > t_c$, the metrics are independent of $|\boldsymbol{t}|_{max}$, which can be demonstrated from the results in Figure S6. The reason is that the valuable information contained in autocorrelations is compressed to the first several PCs, while the lower-ranked PCs containing a large amount of redundant information are forcibly truncated, resulting in the invariance of the low-dimensional features. Therefore, the quantitative metrics based on PCA have the advantage of reliability and robustness in measuring the similarity of physical features of microstructures.

In summary, the differences between our metrics and D_s show up in three ways: (1) In form, D_s reflects the distance between two selected two-point statistics, while our metrics reflect the discrepancy between the low-dimensional PC features. (2) In physical meaning, D_s can only reflect the average degree of similarity in morphology pattern of the two-point statistics, while the rigorous measurement of the distance in physical features (phase volume fraction, average grain size, etc.) of the microstructures can also be addressed by our metrics. (3) In stability, D_s is strongly affected by the dimensionality of the two-point statistics, while our metrics are only related to the microstructures.

Advantages and limitations of the method of optimizing microstructural information

Optimization of microstructural information in this study includes two aspects: construction of RVE set and removement of redundancy. There are two premises for an RVE set: (1) the size of members is large enough to ensure statistical homogeneity in the spatial distribution of structural features that can be mapped to macroscopic mechanical properties; and (2) the dispersion of structural features in the RVE set should match the entire material sample[28]. The microstructures contained in the RVE set are independent of their size and location in the samples [Figure 4 and 5], which meets the first condition. The RVE set absorbs enough structural features that its average autocorrelations converge in PCA space [Figure 6], indicating the second condition has been met. In addition, the method was successfully applied to the datasets of experimental ferrite steels, dendrite solidification of Al-Cu alloys simulated by PFM and experimental Ni-Fe-based superalloys collected in the literature, and SP linkage with high precision was established by Ridge regression. These impressive results demonstrate the advantages of scientific nature, reliability, and universality.

The developed method is an improved version of the average approximation method to construct an RVE set, which is also not readily applicable to the samples with microstructure gradients, for instance, some additively manufactured samples with coarse columnar grains where the "average grain size" characteristic is meaningless[64], the samples with high inhomogeneity in size or distribution of the thermodynamic phases where the average treatment loses the local variation nature of the structure[65], and so on. A rough solution to obtain the statistical information of the overall sample from the SVEs with local gradients is reserving all the original two-point statistics of the SVEs. Nevertheless, dimensional disasters are beyond the scope of conventional dimensionality reduction algorithms, such as PCA. If one insists on extracting two-point statistics of the sample by ensemble averaging the statistics from multiple SVEs, the long-range correlations may be missed. In other words, the SVEs size used to construct an RVE set must exceed the coherence length of the microstructure when the long-range order plays a significant role in the physics of the system[28]. Therefore, the construction of RVE for the samples with microstructure gradients or inhomogeneity remains a challenging task, which is one of the active areas that we will investigate in the future.

Another interesting topic in this study is the tolerance factor ϵ that is used to determine the coherence length t_c, as expressed by Equation (1). From the Equation, one can see that, once t_c is determined, the values of autocorrelations with $|\boldsymbol{t}|$ greater than t_c will fluctuate within a short interval $[\left({}^{h}f^{j}\right)^2 - \epsilon, \left({}^{h}f^{j}\right)^2 + \epsilon]$. As for a microstructure with time dependence and strongly coupled and long-range phenomena such as diffusion, t_c must change over time, and so does the size of RVE[66]. However, the interval length 2ϵ above is a scalar related to twice the margin of the error limitation between autocorrelations and volume fraction squared when $|\boldsymbol{t}|$ is larger than t_c, which is independent of time. In other words, evolution time and long-range phenomena may have a significant effect on t_c but not ϵ. Based on the aging microstructures of ferrite steels and the creep microstructures of the collected Ni-Fe-based superalloys, we confirmed the critical threshold of ϵ to be 0.005 when the main statistical information remained. To verify the inference that ϵ may be also suitable for the evolving microstructures with long-range diffusion, a case about the coarsening process of the polydisperse particle during evolution was used. Our previous work investigated the effects of two characters of the particle cluster, i.e., particle number (N_c) and particle density in a cluster, on the kinetics of transient coarsening[67]. The microstructures in Figure 3 by Wang *et al.* were used to analyze the problem above[67].

Figure S11 provides these microstructures (with four groups of different combinations of N_c and ρ_c), and the variation of pair correlation function $P_{|\boldsymbol{t}|}$ with different $|\boldsymbol{t}|_{max}$; the detailed calculation process is also illustrated in Figure S11. For each group of the parameters, t_c was determined by locating the minimal $|\boldsymbol{t}|_{max}$ when $P_{|\boldsymbol{t}|}$ curves appear platform. Interestingly, as shown in Figure S12, even if t_c changes with different degrees over evolving time, $\left|P_{|\boldsymbol{t}|\geq t_c} - P_0^2\right|$, i.e., ϵ, is still at a low level (< 0.005), indicating t_c is affected by the solute diffusion during evolving and ϵ is indeed independent of evolving time or long-range diffusion. Additionally, ϵ of 0.005 may be a generalized metric to determine t_c of a microstructure, which can be demonstrated from the results shown in Figure 10, Figure S9, and the gray shadow area in Figure S11.

Application prospects and limitations for the proposed scheme

In the practical application of the proposed scheme, flexible feature selection is allowed, such as the addition of other necessary factors in addition to the low-dimensional PC features of microstructures. The factors here can be directly measured from the material samples or filtered by feature engineering. The factors that may be important to the yield strength (grain size, precipitated phase, dislocation, *etc*.) were indeed ignored in our study, resulting in a seemingly "capped" predictability of the final model even with the optimized parameter selection, i.e., R^2 value of $< 90\%$. Strictly speaking, these factors should be incorporated into our scheme to produce a more scientific and robust SP linkage. However, one main contribution of this research is to provide a practical computing strategy for constructing an RVE set. The size and number of small sub-domains in the microstructures is the final optimization objective. Our scheme successfully addressed this goal, although the established SP linkage could be further improved by considering more factors. If readers attempt to use the scheme to predict the mechanical properties of interest related to microstructures, more characterization and/or measurements conducted with expert knowledges are suggested to obtain sufficient inputs of the ML model. It is important to note that these supplementary factors and low-dimensional PC features can be combined to train an ML model, as done in this study. The only effort required is to increase the number of input features.

A major strength of the proposed scheme comes from its ability to extract reliable low-dimensional features by optimizing structural information in an RVE set. Using the features, one can place the microstructures into correct classes by flexibly combining supervised or semi-supervised ML algorithms to study the relationship between structural features and resulting properties for a special material system such as Ag-Al-Cu ternary eutectic alloys or superalloys with multiple strengthening patterns[40,43,68]. Therefore, the scheme can accelerate and improve the procedure of microstructure classification. In addition, our previous study proved that introducing two-point statistical information on microstructures can enhance PSP linkages, which may be further improved by the scheme in this study[7]. We believe that it is competent to predict mechanical properties for most material systems, especially in the case of long-term service in a harsh experimental environment[31–33]. Unfortunately, to our best knowledge, there is no effort to apply two-point statistical information to realize the goal of alloy design. Molkeri *et al*. proved that explicit incorporation of microstructure knowledge in the materials design framework can significantly enhance the materials optimization process[8], and we previously developed an iterative strategy to search ultra-strength martensitic stainless steels in a global-oriented manner[48]. These two studies provide confidence that our scheme [Figure 3], combined with the previously proposed iteration strategy, can be applied to rapidly discover new alloys in experiments. In conclusion, there is a broad application prospect of our scheme in microstructure classification, property prediction, and reverse engineering for designing new materials.

CONCLUSIONS

We propose a novel scheme to construct SP linkage by optimizing microstructure information, which is achieved via employing RI2SS, PCA, and Ridge regression. A small experimental dataset for ferrite steels was created. We designed reliable and robust distance metrics, $\overline{d^{nor}_{centroid}}$, $\overline{d^{nor}_{target}}$, and $d_{\left|\overline{\alpha^{j}_{S^*}} - \overline{\alpha^{j-1}_{S^*}}\right|}$, to quantify the optimal size

(S_{20X}) and number (6) of members in the RVE set. While constructing the set, an innovative method called recursive addition was developed. The primary SP linkage keeps a high accuracy ($R^2 = 0.8680$, $MAE < 10$MPa). After removing redundant information in the autocorrelations, the accuracy was obviously improved by 37.2% ($R^2 = 0.8941$, $MAE < 6.28$MPa). As another contribution of this work, the threshold of tolerance factor ϵ that determines the coherence length in a microstructure was confirmed to be 0.005. Finally, the scientific nature, reliability, and universality of this scheme were proved by performing experiments on two other datasets (dendrite solidification data of Al-Cu alloys simulated by PFM and experimental Ni-Fe-based superalloys data collected in the literature). More importantly, broad application prospects in microstructure classification, property prediction, and alloy design are expected for the scheme.

DECLARATIONS

Authors' contributions

Conception and design of the study: Hu X, Li J, Wang J
Data analysis, visualization, and interpretation: Hu X, Zhao J, Wang J
Method design and modeling: Hu X, Wang J
Materials preparation: Hu X, Chen y, Zhao j, Wangm Y, Wu Q
Writing: Hu X, Li J, Wang Z, Wang J
Review and editing: Li J, Wang Z, Wang J
Resources, supervision, and project administration: Li J, Wang J

Availability of data and materials

Supplementary materials are available from the Journal of Materials Informatics or from the authors.

Financial support and sponsorship

This work was supported by the National Natural Science Foundation of China (Grant No. 51871183 and 51874245). The authors also thank the High-Performance Computing Center of Northwestern Polytechnical University, China, for the computer time and facilities.

Conflicts of interest

The authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

©The Author(s) 2022.

Supplementary materials

Supplementary materials

REFERENCES

1. Rajan K. Materials informatics. *Materials Today* 2005;8:38-45. DOI
2. Hart GLW, Mueller T, Toher C, Curtarolo S. Machine learning for alloys. *Nat Rev Mater* 2021;6:730-55. DOI
3. Agrawal A, Choudhary A. Perspective: materials informatics and big data: realization of the "fourth paradigm" of science in materials science. *APL Mater* 2016;4:053208. DOI

4. Kalidindi SR, Niezgoda SR, Salem AA. Microstructure informatics using higher-order statistics and efficient data-mining protocols. *JOM* 2011;63:34-41. DOI
5. Kalidindi SR, De Graef M. Materials data science: current status and future outlook. *Annu Rev Mater Res* 2015;45:171-93. DOI
6. Arróyave R, Mcdowell DL. Systems approaches to materials design: past, present, and future. *Annu Rev Mater Res* 2019;49:103-26. DOI
7. Hu X, Li J, Wang Z, Wang J. A microstructure-informatic strategy for Vickers hardness forecast of austenitic steels from experimental data. *Materials & Design* 2021;201:109497. DOI
8. Molkeri A, Khatamsaz D, Couperthwaite R, et al. On the importance of microstructure information in materials design: PSP vs PP. *Acta Materialia* 2022;223:117471. DOI
9. Khatavkar N, Swetlana S, Singh AK. Accelerated prediction of Vickers hardness of Co- and Ni-based superalloys from microstructure and composition using advanced image processing techniques and machine learning. *Acta Materialia* 2020;196:295-303. DOI
10. Shen C, Wang C, Wei X, Li Y, van der Zwaag S, Xu W. Physical metallurgy-guided machine learning and artificial intelligent design of ultrahigh-strength stainless steel. *Acta Materialia* 2019;179:201-14. DOI
11. Shin D, Yamamoto Y, Brady M, Lee S, Haynes J. Modern data analytics approach to predict creep of high-temperature alloys. *Acta Materialia* 2019;168:321-30. DOI
12. Peng J, Yamamoto Y, Hawk JA, Lara-curzio E, Shin D. Coupling physics in machine learning to predict properties of high-temperatures alloys. *npj Comput Mater* 2020;6. DOI
13. Hu X, Wang J, Wang Y, et al. Two-way design of alloys for advanced ultra supercritical plants based on machine learning. *Computational Materials Science* 2018;155:331-9. DOI
14. Niezgoda SR, Kanjarla AK, Kalidindi SR. Novel microstructure quantification framework for databasing, visualization, and analysis of microstructure data. *Integr Mater Manuf Innov* 2013;2:54-80. DOI
15. Sangid MD, Ravi P, Prithivirajan V, Miller NA, Kenesei P, Park J. ICME approach to determining critical pore size of IN718 produced by selective laser melting. *JOM* 2020;72:465-74. DOI
16. Zinovieva O, Romanova V, Balokhonov R. Effects of scanning pattern on the grain structure and elastic properties of additively manufactured 316L austenitic stainless steel. *Materials Science and Engineering: A* 2022;832:142447. DOI
17. Popova E, Rodgers TM, Gong X, Cecen A, Madison JD, Kalidindi SR. Process-structure linkages using a data science approach: application to simulated additive manufacturing data. *Integr Mater Manuf Innov* 2017;6:54-68. DOI PubMed PMC
18. Yabansu YC, Rehn V, Hötzer J, Nestler B, Kalidindi SR. Application of Gaussian process autoregressive models for capturing the time evolution of microstructure statistics from phase-field simulations for sintering of polycrystalline ceramics. *Modelling Simul Mater Sci Eng* 2019;27:084006. DOI
19. Latypov MI, Kühbach M, Beyerlein IJ, et al. Application of chord length distributions and principal component analysis for quantification and representation of diverse polycrystalline microstructures. *Materials Characterization* 2018;145:671-85. DOI
20. Turner DM, Niezgoda SR, Kalidindi SR. Efficient computation of the angularly resolved chord length distributions and lineal path functions in large microstructure datasets. *Modelling Simul Mater Sci Eng* 2016;24:075002. DOI
21. Yucel B, Yucel S, Ray A, Duprez L, Kalidindi SR. Mining the correlations between optical micrographs and mechanical properties of cold-rolled HSLA steels using machine learning approaches. *Integr Mater Manuf Innov* 2020;9:240-56. DOI
22. Niezgoda S, Fullwood D, Kalidindi S. Delineation of the space of 2-point correlations in a composite material system. *Acta Materialia* 2008;56:5285-92. DOI
23. Cecen A, Yabansu YC, Kalidindi SR. A new framework for rotationally invariant two-point spatial correlations in microstructure datasets. *Acta Materialia* 2018;158:53-64. DOI
24. Fullwood DT, Niezgoda SR, Adams BL, Kalidindi SR. Microstructure sensitive design for performance optimization. *Progress in Materials Science* 2010;55:477-562. DOI
25. Brough DB, Wheeler D, Warren JA, Kalidindi SR. Microstructure-based knowledge systems for capturing process-structure evolution linkages. *Current Opinion in Solid State and Materials Science* 2017;21:129-40. PubMed PMC
26. Fullwood DT, Niezgoda SR, Kalidindi SR. Microstructure reconstructions from 2-point statistics using phase-recovery algorithms. *Acta Materialia* 2008;56:942-8. DOI
27. Bostanabad R, Zhang Y, Li X, et al. Computational microstructure characterization and reconstruction: review of the state-of-the-art techniques. *Progress in Materials Science* 2018;95:1-41. DOI
28. Niezgoda SR, Turner DM, Fullwood DT, Kalidindi SR. Optimized structure based representative volume element sets reflecting the ensemble-averaged 2-point statistics. *Acta Materialia* 2010;58:4432-45. DOI
29. Khosravani A, Cecen A, Kalidindi SR. Development of high throughput assays for establishing process-structure-property linkages in multiphase polycrystalline metals: Application to dual-phase steels. *Acta Materialia* 2017;123:55-69. DOI
30. Steinmetz P, Yabansu YC, Hötzer J, Jainta M, Nestler B, Kalidindi SR. Analytics for microstructure datasets produced by phase-field simulations. *Acta Materialia* 2016;103:192-203. DOI
31. Yabansu YC, Iskakov A, Kapustina A, Rajagopalan S, Kalidindi SR. Application of Gaussian process regression models for capturing the evolution of microstructure statistics in aging of nickel-based superalloys. *Acta Materialia* 2019;178:45-58. DOI
32. Iskakov A, Yabansu YC, Rajagopalan S, Kapustina A, Kalidindi SR. 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 Materialia* 2018;144:758-67. DOI
33. Gorgannejad S, Reisi Gahrooei M, Paynabar K, Neu R. Quantitative prediction of the aged state of Ni-base superalloys using PCA and

tensor regression. *Acta Materialia* 2019;165:259-69. DOI
34. Yabansu YC, Steinmetz P, Hötzer J, Kalidindi SR, Nestler B. Extraction of reduced-order process-structure linkages from phase-field simulations. *Acta Materialia* 2017;124:182-94. DOI
35. Tewari A, Gokhale A, Spowart J, Miracle D. Quantitative characterization of spatial clustering in three-dimensional microstructures using two-point correlation functions. *Acta Materialia* 2004;52:307-19. DOI
36. Jung J, Yoon JI, Park HK, Kim JY, Kim HS. Bayesian approach in predicting mechanical properties of materials: application to dual phase steels. *Materials Science and Engineering: A* 2019;743:382-90. DOI
37. Gupta A, Cecen A, Goyal S, Singh AK, Kalidindi SR. Structure-property linkages using a data science approach: application to a non-metallic inclusion/steel composite system. *Acta Materialia* 2015;91:239-54. DOI
38. Bro R, Smilde AK. Principal component analysis. *Anal Methods* 2014;6:2812-31. DOI
39. Fast T, Wodo O, Ganapathysubramanian B, Kalidindi SR. Microstructure taxonomy based on spatial correlations: application to microstructure coarsening. *Acta Materialia* 2016;108:176-85. DOI
40. Kunselman C, Attari V, Mcclenny L, Braga-neto U, Arroyave R. Semi-supervised learning approaches to class assignment in ambiguous microstructures. *Acta Materialia* 2020;188:49-62. DOI
41. Kitahara AR, Holm EA. Microstructure cluster analysis with transfer learning and unsupervised learning. *Integr Mater Manuf Innov* 2018;7:148-56. DOI
42. Liu Q, Wu H, Paul MJ, et al. Machine-learning assisted laser powder bed fusion process optimization for AlSi10Mg: new microstructure description indices and fracture mechanisms. *Acta Materialia* 2020;201:316-28. DOI
43. Choudhury A, Yabansu YC, Kalidindi SR, Dennstedt A. Quantification and classification of microstructures in ternary eutectic alloys using 2-point spatial correlations and principal component analyses. *Acta Materialia* 2016;110:131-41. DOI
44. Wu Q, Wang Z, Hu X, et al. Uncovering the eutectics design by machine learning in the Al-Co-Cr-Fe-Ni high entropy system. *Acta Materialia* 2020;182:278-86. DOI
45. Zheng T, Hu X, He F, et al. Tailoring nanoprecipitates for ultra-strong high-entropy alloys via machine learning and prestrain aging. *Journal of Materials Science & Technology* 2021;69:156-67. DOI
46. Nelson J, Sanvito S. Predicting the Curie temperature of ferromagnets using machine learning. *Phys Rev Materials* 2019;3. DOI
47. Mukhamedov BO, Karavaev KV, Abrikosov IA. Machine learning prediction of thermodynamic and mechanical properties of multicomponent Fe-Cr-based alloys. *Phys Rev Materials* 2021;5. DOI
48. Hu X, Zhao J, Li J, Wang Z, Chen Y, Wang J. Global-oriented strategy for searching ultrastrength martensitic stainless steels. *Advcd Theory and Sims* 2022;5:2100411. DOI
49. Masuyama F. Advances in physical metallurgy and processing of steels. History of power plants and progress in heat resistant steels. *ISIJ International* 2001;41:612-25. DOI
50. Konadu DS, Pistorius PGH. Investigation of formation of precipitates and solidification temperatures of ferritic stainless steels using differential scanning calorimetry and Thermo-Calc simulation. *Sādhanā* 2021;46. DOI
51. Rojas D, Garcia J, Prat O, Sauthoff G, Kaysser-pyzalla A. 9%Cr heat resistant steels: alloy design, microstructure evolution and creep response at 650°C. *Materials Science and Engineering: A* 2011;528:5164-76. DOI
52. Knezevic V, Sauthoff G, Vilk J, et al. Martensitic/Ferritic super heat-resistant 650.DEG.C. steels. Design and testing of model alloys. *ISIJ International* 2002;42:1505-14. DOI
53. Li X, Kuang W, Zhang J, Zhou Q, Wang H. Application of the thermodynamic extremal principle to massive transformations in Fe-C alloys. *Metall and Mat Trans A* 2018;49:4484-94. DOI
54. Zhong Z, Gu Y, Yuan Y. Microstructural stability and mechanical properties of a newly developed Ni-Fe-base superalloy. *Materials Science and Engineering: A* 2015;622:101-7. DOI
55. Otsu N. A threshold selection method from gray-level histograms. *IEEE Trans Syst , Man, Cybern* 1979;9:62-6. DOI
56. van der Walt S, Schönberger JL, Nunez-Iglesias J, et al; scikit-image contributors. Scikit-image: image processing in python. *PeerJ* 2014;2:e453. DOI PubMed PMC
57. Marquardt DW, Snee RD. Ridge regression in practice. *The American Statistician* 1975;29:3-20. DOI
58. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Mach Learn Res* 2011;12:2825-28306. Available from: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf [Last accessed on 20 Mar 2022]
59. Li J, Wang Z, Wang Y, Wang J. Phase-field study of competitive dendritic growth of converging grains during directional solidification. *Acta Materialia* 2012;60:1478-93. DOI
60. Guo C, Li J, Yu H, Wang Z, Lin X, Wang J. Branching-induced grain boundary evolution during directional solidification of columnar dendritic grains. *Acta Materialia* 2017;136:148-63. DOI
61. Wang Y, Li J, Yang H, et al. The formation mechanism of special globular surface grain during the solidification of laser surface remelted near β titanium alloys. *Computational Materials Science* 2021;191:110353. DOI
62. Eiken J, Böttger B, Steinbach I. Multiphase-field approach for multicomponent alloys with extrapolation scheme for numerical application. *Phys Rev E Stat Nonlin Soft Matter Phys* 2006;73:066122. DOI PubMed
63. Farzadi A, Do-quang M, Serajzadeh S, Kokabi AH, Amberg G. Phase-field simulation of weld solidification microstructure in an Al-Cu alloy. *Modelling Simul Mater Sci Eng* 2008;16:065005. DOI
64. Wang F, Williams S, Rush M. Morphology investigation on direct current pulsed gas tungsten arc welded additive layer manufactured Ti6Al4V alloy. *Int J Adv Manuf Technol* 2011;57:597-603. DOI
65. Zinovieva O, Zinoviev A, Romanova V, Balokhonov R. Three-dimensional analysis of grain structure and texture of additively manu-

factured 316L austenitic stainless steel. *Additive Manufacturing* 2020;36:101521. DOI

66. Schwarze C, Darvishi Kamachali R, Kühbach M, et al. Computationally efficient phase-field simulation studies using RVE sampling and statistical analysis. *Computational Materials Science* 2018;147:204-16. DOI
67. 67. Wang Y, Li J, Zhang L, Wang Z, Wang J. Phase-field study on the effect of initial particle aggregation on the transient coarsening behaviors. *Modelling Simul Mater Sci Eng* 2020;28:075007. DOI
68. 68. Liu Y, Wu J, Wang Z, et al. Predicting creep rupture life of Ni-based single crystal superalloys using divide-and-conquer approach based machine learning. *Acta Materialia* 2020;195:454-67. DOI