Automated Analysis of Continuum Fields from Atomistic Simulations Using Statistical Machine Learning

Aruna Prakash* and Stefan Sandfeld

Atomistic simulations of the molecular dynamics/statics kind are regularly used to study small-scale plasticity. Contemporary simulations are performed with tens to hundreds of millions of atoms, with snapshots of these configurations written out at regular intervals for further analysis. Continuum scale constitutive models for material behavior can benefit from information on the atomic scale, in particular in terms of the deformation mechanisms, the accommodation of the total strain, and partitioning of stress and strain fields in individual grains. Herein, a methodology is developed using statistical data mining and machine learning algorithms to automate the analysis of continuum field variables in atomistic simulations. Three important field variables are focused on: total strain, elastic strain, and microrotation. The results show that the elastic strain in individual grains exhibits a unimodal lognormal distribution, while the total strain and microrotation fields evidence a multimodal distribution. The peaks in the distribution of total strain are identified with a Gaussian mixture model and methods to circumvent overfitting problems are presented. Subsequently, the identified peaks are evaluated in terms of deformation mechanisms in a grain, which, e.g., helps to quantify the strain for which individual deformation mechanisms are responsible. The overall statistics of the distributions over all grains are an important input for higher scale models, which ultimately also helps to be able to quantitatively discuss the implications for information transfer to phenomenological models.

A. Prakash, S. Sandfeld
Micromechanical Materials Modelling Group (MiMM)
Institute of Mechancis and Fluid Dynamics
Technical University Bergakademie Freiberg
Lampadiusstraße 4, 09599 Freiberg, Germany
E-mail: arun.prakash@imfd.tu-freiberg.de

S. Sandfeld
Institute for Advanced Simulation—IAS-9
Materials Data Science and Informatics
Forschungszentrum Juelich GmbH
52425 Zurich, Germany

S. Sandfeld
Chair of Materials Data Science and Materials Informatics
Faculty 5—Georesources and Materials Engineering
RWTH Aachen University
52056 Aachen, Germany

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/adem.202200574.

1. Introduction

Over the last decade, data science and informatics (DSI) has evolved as the fourth paradigm of scientific research,[1] in addition to the traditional paradigms of experiments/empirical reasoning, theory/modeling, and computation/simulation, and has shown great potential for accelerated materials development.[2,3] A characteristic feature of approaches and predictive methods from DSI is that they focus strongly on the data itself while still allowing to consider physical domain knowledge.[4] This essentially makes it particularly attractive for increasing the synergy between, e.g., experiments and simulations. Challenges remain, nonetheless, concerning the availability of data—usually summarized through the "four Vs" (volume, variety, veracity, and velocity)—because algorithms and methods in DSI typically require well-curated and sufficiently large datasets.[5] In recent years, high-throughput experiments with improved imaging techniques as well as large-scale computations, e.g., atomistic simulations, with high-performance computing resources have helped significantly in this regard.[4]

Atomistic simulations have now become an invaluable tool in the field of computational materials science and have been particularly useful in advancing our knowledge on the mechanical behavior of materials.[6] Such simulations are regularly used to study defect–defect interactions,[7–10] elastoplasticity,[11–13] fracture,[14,15] irradiation,[16–19] and other phenomena in crystalline materials. Large-scale computations have allowed researchers to study fundamental processes in nanoscale objects like thin films,[20,21] nanoparticles,[22–24] and nanowires,[10,25,26] and have, furthermore, facilitated experimentally informed simulations,[27–29] resulting in an improved fundamental understanding of small-scale plasticity.

Studies with atomistic simulations on nanocrystalline materials—the focus of the current work—have been successful in elucidating the deformation mechanisms and the role of grain boundaries (GBs) in polycrystalline materials with grain sizes below 100 nm.[30,31] Such studies have demonstrated the role of GBs as sources and sinks for dislocations, due to the lack of nucleation sources like Frank–Read or spiral sources in individual grains.[32,33] Furthermore, the inverse Hall–Petch effect characterized by a decrease in strength with decreasing grain size

below roughly 10–20 nm is attributed to a change in deformation mechanism from that dominated by intragrain dislocations to grain boundary mediated deformation.[34]

Efforts have been made to link atomistic data with continuum scale data and simulations.[25,35–38] This is particularly important for the development of robust and reliable constitutive models that are able to consider emergent properties from the nanoscale, where, e.g., plasticity is governed by a paucity of dislocation sources and an increased fraction of grain boundaries, interfaces, and surfaces. Continuum scale modeling of complex material phenomena at the nanoscale hinges not only on the availability of numerical values of parameters for constitutive models, but also on information regarding the relative contribution of individual mechanisms to the total deformation, and statistical data pertaining to the distribution of fields like strain, stress, texture, etc. Obtaining such information is generally tedious and time-consuming due to a lack of automation.[13,39,40]

Data mining and the use of machine learning algorithms have become extremely popular in materials science covering many different problem classes ranging from analysis of microscopy data[41–43] to design of alloys and metamaterials, accelerated materials discovery, and machine learning-guided theory development.[44–48] Such machine learning methods can also help overcome the problem of obtaining the aforementioned information in an automated manner and facilitate knowledge transfer between the atomistic and continuum scales.[47–50] Additionally, one of the challenges in machine learning of "never enough data" is easily overcome because every atom is essentially a data point—i.e., fields of interest, like total and elastic strain, are calculated as properties of individual atoms—and typical simulations involve millions to billions of atoms. Mining such data will help understand the complex relationships between the emerging local fields—such as strain, stress, and texture—with the macroscopic response, and support in the formulation of microstructure–property relationships.[51,52]

In this work, we develop a methodology for automated analysis and visualization of continuum fields like strains and rotations in large-scale atomistic simulations. In particular, we use data mining and statistical machine learning algorithms to extract key features from distributions of total strain, elastic strain, and rotation in individual grains. Generally, machine learning tasks are classified into two broad categories:[53] descriptive and predictive. The former approach is of explanatory nature and aims to identify patterns in the underlying data based on correlations and trends, while the latter approach is used to predict and foresee events induced by certain known factors.

Herein, the focus in the development of this methodology is primarily of descriptive nature. We develop an understanding of unimodal distributions by identifying the mathematical form of the distribution. For multimodal distributions, the values of individual peaks are identified using a Gaussian mixture model (GMM), wherein methods to circumvent overfitting problems are presented. The individual peaks are then correlated with the deformation mechanisms observed in a grain. Finally, approaches to make the methodology a predictive one are briefly discussed.

2. Details on the Data Used

The data used in the current work are sourced from a tensile simulation of a nanocrystalline thin film sample. The initial structure, with a mean grain size of 15 nm, is generated by means of a constrained Voronoi tessellation (CVT)[13,54] so as to reduce nonequilibrium junctions.[55] The initial dimensions of the thin film are 180 nm × 120 nm × 15 nm (see **Figure 1**a). The atomistic sample subsequently generated using the open source toolbox nanoSCULPT[56] contains ≈19 Mio atoms. The interatomic forces are modeled with an embedded atom method (EAM) potential for Al[57] using a stable time increment of 1 fs. The structure is relaxed using the fire algorithm in standard molecular statics simulations up to a force norm of 10^{-7} eV Å^{-1}, and subsequently equilibrated at 300 K for 20 ps. All simulations are performed with the atomistic simulation code IMD.[58] Careful initial tests were conducted with the interatomic potential to obtain numerical values of 45.8 (inv_tau_eta in IMD) for the thermostat and 45.0 (inv_tau_xi in IMD) for the barostat. Subsequently, constant pressure simulations are performed at 300 K using a Nosé–Hoover thermobarostat to ensure close to zero global stresses in directions normal to the simulation box.

The relaxed and equilibrated structures are then subjected to uniaxial tension at a constant strain rate of $\dot{\varepsilon} = 10^9$ s^{-1} at 300 K in the NPT ensemble. Uniaxial strain is modeled by continuously scaling the atomic coordinates and the box length along the y direction at a fixed rate while simultaneously allowing for contraction along the x direction using a Nosé–Hoover thermobarostat. Periodic boundary conditions are imposed along the x and y directions; free boundaries exist in the z direction along the thickness of the film.

Snapshots are written out every 0.05% strain increment (time increments of 5 ps). Each snapshot is a separate file in a simple ASCII file format containing information on the position and index of each atom. The index has no physical meaning and has no bearing on the simulations itself. It is merely used to identify each atom uniquely and can hence be used to map an atom to a particular grain in the initial configuration.

The so-obtained snapshot files are subsequently analyzed using the open-source visualization tool Ovito.[59] Defect structures are identified using the common neighbor analysis (CNA) modifier.[60] Local atomic strains are determined using the atomic strain modifier, which calculates the deformation gradient tensor $\hat{\mathbf{F}}$ using the displacement of individual atoms, and subsequently, the Green–Lagrange strain tensor $\hat{\mathbf{E}}$ as follows

$$\hat{\mathbf{E}} = \frac{1}{2}(\hat{\mathbf{F}}^{\mathrm{T}}\hat{\mathbf{F}} - \mathbf{1}) \tag{1}$$

This atomic strain tensor is then averaged over the neighborhood of each atom using a cutoff (or smoothing) radius of 1 nm to obtain an averaged total strain tensor per atom

$$\mathbf{E} = \frac{\sum_i^N \hat{\mathbf{E}}_i \hat{V}_i}{\sum_i^N \hat{V}_i} \tag{2}$$

where $\hat{V}$ denotes the volume of each atom as determined by a Voronoi construction and $\mathbf{E}$ provides a measure of the total strain. The choice of 1 nm for the smoothing radius is based on
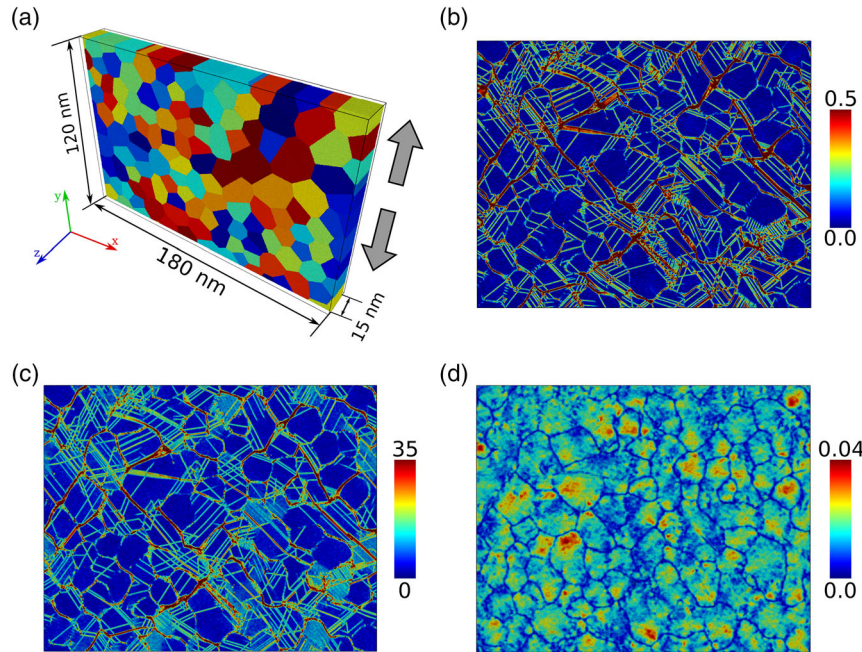
Figure 1. Atomistic sample used as data source in the current work: a) thin film sample in the undeformed configuration along with sample dimensions and global directions. The loading direction (global y direction) is shown via gray arrows. Atoms are colored according to their grain number; b–d) Map of E_{vm}, $\Delta\psi$, and E_{vm}^{el}, respectively, in individual grains in the deformed state after 10% applied strain. The rotation angle in (c) is in degrees. For further statistical analysis, the distributions of these three quantities are analyzed over the domain of an individual grain.

previous studies on stress distribution in γ/γ' microstructures.[29] In the current study, no significant differences were observed between smoothing radii of 0.5 and 1.0 nm. Essentially, the value of the smoothing radii must include enough neighbor shells to remove any noise in the field of interest. Too large a value of the smoothing radius smears out inhomogeneities, and in the limiting case of smoothing radius larger than the sample dimensions, the average response, i.e., the macroscopic strain tensor, is to be expected.

The von Mises equivalent total strain E_{vm} is then computed using the averaged strain tensor as follows[61]

$$E_{vm} = \sqrt{E_{xy}^2 + E_{yz}^2 + E_{zx}^2 + \frac{1}{6}\left[(E_{xx} - E_{yy})^2 + (E_{xx} - E_{zz})^2 + (E_{yy} - E_{zz})^2\right]}$$

$$(3)$$

As $\hat{F}$ is already known, we may perform the polar decomposition to obtain the rotation tensor $\hat{R}$ and the stretch tensor $\hat{U}$ as $\hat{F} = \hat{R}\,\hat{U}$. The so-computed rotation tensor is then averaged over the a shell of next-nearest neighbors to obtain the averaged rotation tensor R, and is expressed as a quaternion (q, q_0). We now compute the incremental angle of rotation $\Delta\psi$, called microrotation, as

$$\Delta\psi = 2 \cdot \cos^{-1}(q_0) \qquad (4)$$

which provides a measure of the rotation experienced by an atom in the deformed sample, with respect to the initial configuration.

The local elastic strain $\hat{E}^{el}$ is computed using the elastic strain modifier in Ovito and averaged with an identical procedure as for the Green–Lagrange strain tensor E. Essentially, Ovito computes the local elastic deformation gradient $\hat{F}^e$ at each atom by taking into account the positions of the atoms in the local neighborhood and comparing it to that of an ideal crystal structure. A value of 4.07 Å is used as the lattice parameter for the ideal crystal structure. For close packed structures, 12 nearest neighbors are used for the computation. Any distortion in comparison to the ideal crystal is then captured by $\hat{F}^e$, using which the local elastic strain $\hat{E}^{el}$ is computed. Subsequently, a von Mises equivalent elastic strain is computed using an equivalent relation as in Equation (3). For more details on the computation of the elastic strain, the reader is referred to user manual of Ovito.[62]

The analyzed snapshots are now written out as individual files in the IMD ASCII file format. Each file contains the original snapshot data (position and index) appended with the a posteriori calculated data via Ovito. Each column represents the distribution of a field quantity evaluated at discrete points given by the atomic coordinates. For subsequent analysis, we restrict ourselves to three columns: ShearStrain (von Mises equivalent total strain E_{vm}), DeltaRot (angle of rotation $\Delta\psi$), and ElasStrVM (von Mises equivalent elastic strain E_{vm}^{el}). Figure 1b–d shows the distribution of the three field quantities on a slice of the thin film after 10% applied strain.

For further statistical analysis, these field quantities are analyzed over the domain of an individual grain. Only those atoms identified as belonging to a perfect crystal symmetry, i.e., fcc, hcp, etc., are used. All atoms are provided with a unique grain number to identify the grain they belong to. This is done by using the index of each atom and mapping it to the grain they belong to

in the initial configuration. As a result, the definition of the domain of the grain in the deformed configuration is identical to that in the initial configuration.

Figure 2 shows the distribution of the three field quantities E_{vm}^{el}, E_{vm}, and $\Delta\psi$ in individual grains as 1D heat maps. Representation of distributions as heat maps is an alternative to the traditional methods of representation as histograms and kernel density estimates (KDE). The latter methods require significant amount of space for plotting the data of all grains, which makes comparison of distributions over multiple grains a cumbersome operation. For instance, in the case of the current sample, which corresponds to a particular snapshot at 10% strain from a single simulation, we obtain 366 plots for the three field distributions in 122 grains. Plotting as heatmaps ensures minimal space usage for all the plots—while qualitatively keeping all the details—and facilitates easy comparison of distributions. Nonetheless, for applications that go beyond mere visualization, e.g., determination of the mathematical form of the distribution or a correlation of the distribution with deformation mechanisms in the grain, representation of distributions as histograms or KDEs becomes inevitable.

In general, comparison of plots of multiple grains demands that the set of units used on the ordinate and abscissa axes be identical in all plots. As the distributions are significantly inhomogeneous, manual tuning of the plot limits, which is usually done by inspecting the distributions of only a few grains, can result in truncation of outlying peaks in multimodal distributions. On the other hand, merely using the maximum and minimum value of the data array would result in tails of a distribution with frequency values close to zero dominating the plots of the distributions. This is a nontrivial problem and needs to be solved in an objective manner; otherwise the comparability between plots would suffer. Here, we propose a method for a reasonable, automated determination of the plot range, the details of which are provided in the Supporting Information. Using this automated approach, the distributions are plotted as histograms in Figure S2–S4, Supporting Information. The plots show multiple peaks in the distributions of E_{vm} and $\Delta\psi$, evidencing the multimodal nature of the fields. By comparison, the distribution of E_{vm}^{el} is, by and large, unimodal.

The multimodal nature of the distributions of E_{vm} and $\Delta\psi$ can be explained as follows. In nanocrystalline materials, dislocations are nucleated at grain boundaries due to the paucity of intragranular dislocation sources. Such dislocations then traverse through the grain before being absorbed by the opposite grain boundary. As a result, atoms that experience the movement of such dislocations evidence higher strain than those located elsewhere in the grain. The same argument also holds true for the orientation distribution within a single grain. Consequently, the distributions of the total strain and the orientation display a multimodal character. By contrast, all atoms undergo elastic deformation. As a result, the distribution of elastic strain is likely to be unimodal.

3. Methodology

In order to statistically analyze the field distributions in individual grains, we differentiate between unimodal and multimodal

distributions. Our aim is to determine the location of individual peaks in the multimodal distributions of total strain (E_{vm}) and microrotation ($\Delta\psi$) fields. This information is then used to understand the mechanistic source of the inhomogeneity in the distribution. For the case of unimodal distributions, we aim to identify the mathematical form of a function(al) that can statistically represent the unimodal distribution under consideration, over all grains in the polycrystal. At a later stage, such statistics over multiple samples, strain states, and materials can be compared in order to assess the transferability and tractability of such a mathematical representation to higher length-scale simulation frameworks like crystal plasticity or strain gradient models.

3.1. Mathematical Form of Unimodal Distributions

To identify the mathematical function that best describes a unimodal distribution, we first fit an array of functions to the data of individual grains. The stats library implemented in the SciPy package[63] is used to define various functions and to fit them to the given dataset using the maximum likelihood estimation.[64] For the purpose of this work, we only use a subset of all available functions in the stats library. The chosen candidates are those that are likely to reflect the unimodal distribution under consideration. A total of 32 functions are used in the current work (see **Figure 3** and Table A1 for the function names). The reader is referred to the official documentation of the stats library[63] for the mathematical formulation of the individual functions.

Once the parameters of individual functions that result in the best fit of the given distribution (in an individual grain) are identified, an assessment of the function that provides a good description over all grains needs to be made. For this purpose, we compute the sum of squared errors (SSE) of the fits with individual functions with respect to the data of individual grains. A decision on the best fit function is then made by evaluating the statistics of SSE of individual functions over all grains.

3.2. GMM for Multimodal Distributions

We employ a GMM to identify the peaks in a multimodal distribution. GMMs belong to the class of unsupervised machine learning algorithms that are generally used to cluster data points that share certain common characteristics.[64] It is a probabilistic model in which the clustering function is a mixture of several Gaussian functions. As a result, we obtain the probability with which each data point can be assigned to a particular Gaussian function or cluster. GMMs can be deemed as a generalization of the well-known k-means clustering which is a so-called hard clustering method where each point is assigned definitively to a cluster without an uncertainty measure to qualify this association.[64,65] By contrast, GMMs incorporate information on not just the probability with which a data point can be associated with a cluster, but also the covariance structure of the data, in addition to the centers of the individual Gaussians.

Formally, a GMM is defined as the weighted sum of K component Gaussian densities as follows
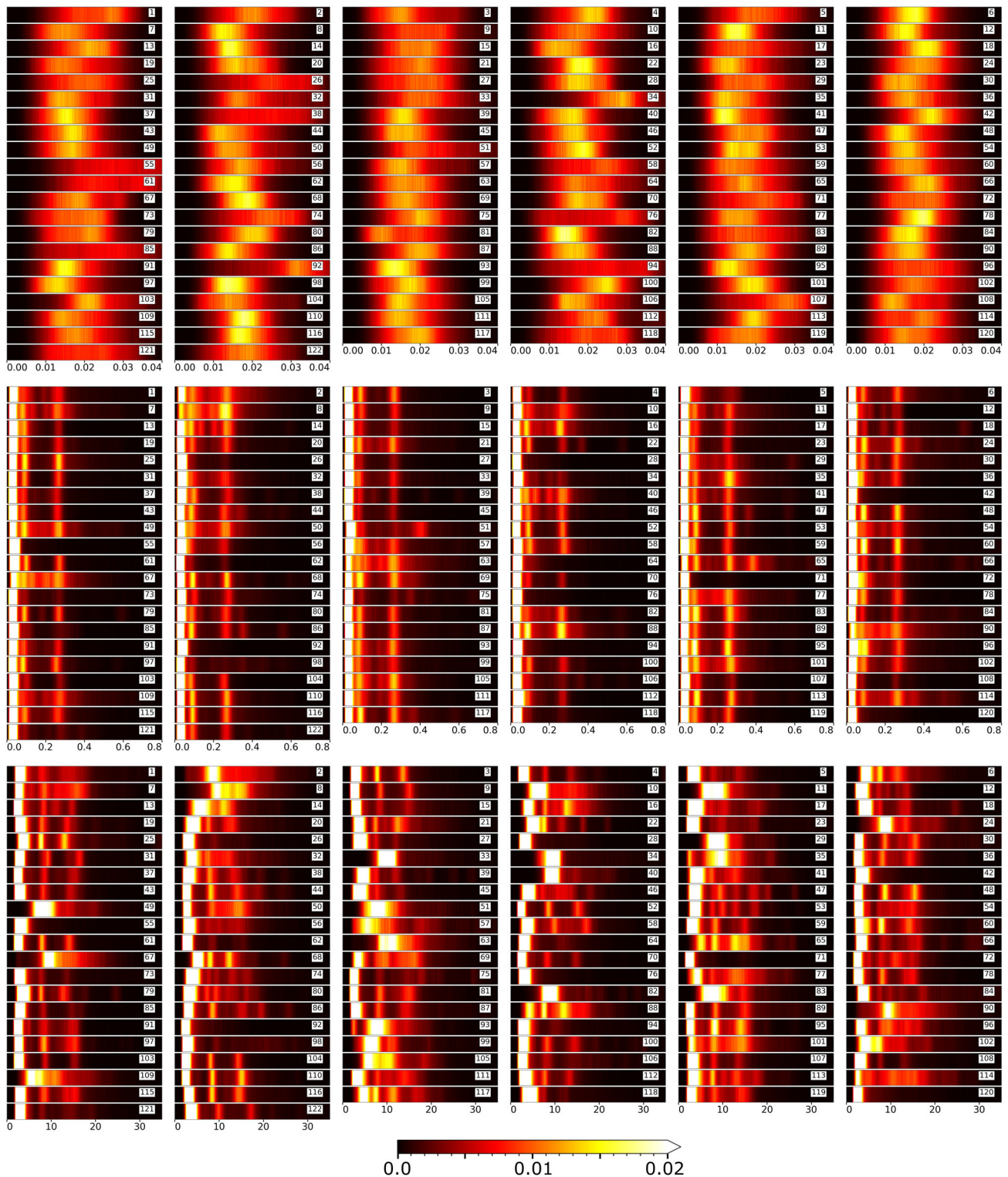
**ADVANCED
SCIENCE NEWS**

www.advancedsciencenews.com

**ADVANCED
ENGINEERING
MATERIALS**

www.aem-journal.com

Figure 2. Distributions of elastic strain E_{vm}^{el} (top row), total strain E_{vm} (middle row), and microrotation $\Delta\psi$ (bottom row) as heat maps, calculated using 200 bins over the corresponding data range. The colormap shows the volume fraction, i.e., the number of atoms in a bin as a fraction of the total number of atoms in the grain. Each horizontal heat map shows the distribution of the corresponding field in a single grain. The grain number is noted in the white box in each heat map.
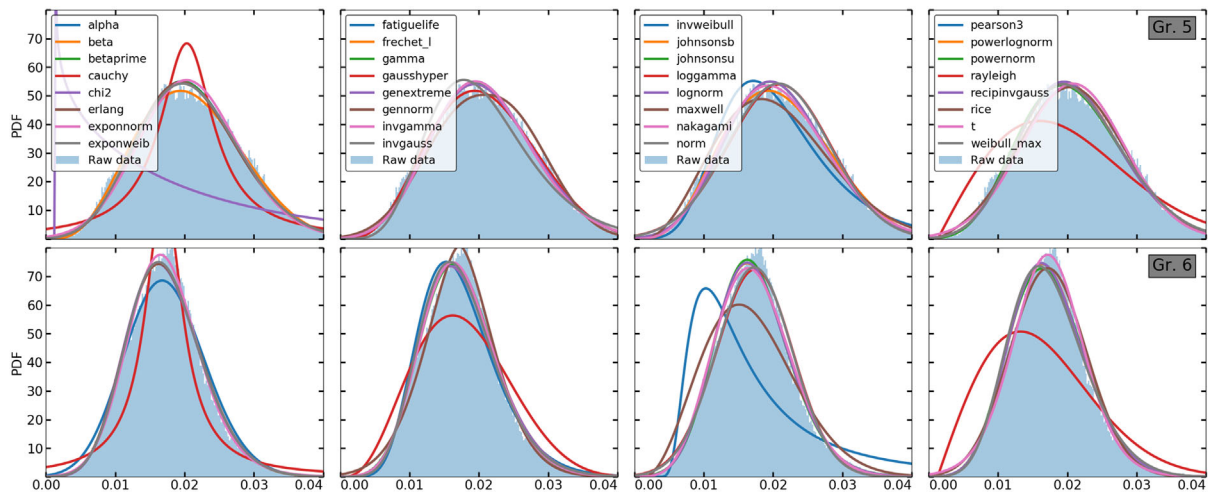
ADVANCED
SCIENCE NEWS
www.advancedsciencenews.com

ADVANCED
ENGINEERING
MATERIALS
www.aem-journal.com

Figure 3. Results of fitting mathematical functions to the distribution of E_{vm}^{el} in two grains 5 and 6. For the exact mathematical expressions of the individual functions, the reader is referred to ref. [63].

$$p(\mathbf{x}) = \sum_{i=1}^{K} w_i \mathcal{N}(\mathbf{x}|\mu_i, \sigma_i) \tag{5}$$

where $p(\mathbf{x})$ is the probability of a data point x, w_i are the weights, and $\mathcal{N}(\mathbf{x}|\mu_i, \sigma_i)$ are the component Gaussian densities, which in D-dimensions are given by

$$\mathcal{N}(\mathbf{x}|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{ -\frac{1}{2}(\mathbf{x} - \mu_i)^T \Sigma_i^{-1}(\mathbf{x} - \mu_i) \right\} \tag{6}$$

with the mean vector μ_i and the covariance matrix Σ_i. The mixture weights w_i are defined as the partition of unity, i.e.

$$\sum_{i=1}^{K} w_i = 1 \tag{7}$$

In the case of a 1D data set, Equation (6) reduces to

$$\mathcal{N}(\mathbf{x}|\mu_i, \sigma_i) = \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{(x - \mu_i)^2}{2\sigma_i^2} \right). \tag{8}$$

For a fixed number of Gaussian components, the parameters μ_i, σ, and w_i can be learned using the expectation-maximization (EM) algorithm. EM is a numerical realization of maximum likelihood estimation and has the convenient property that the algorithm tends toward a local optimum with every iteration.

In the current work, the classical non-Bayesian GMM together with the EM algorithm as implemented in scikit-learn[66] is used. To identify the individual peaks, we fit GMMs with different number of components to the multimodal distribution under consideration, and choose the best fit using certain criteria. The value of individual peaks is then simply the mean of the corresponding Gaussian component. More details on the application of GMM to the atomistic dataset under consideration are provided in Section 4.2.

4. Results

4.1. Mathematical Form of Elastic Strain Distribution

Unimodal distributions, like that of E_{vm}^{el} here, can be evaluated and better understood via the functional form defining the distribution in individual grains. To identify the mathematical function that best fits the distribution of E_{vm}^{el}, we fit various functions to the distribution of E_{vm}^{el} in each grain using a maximum likelihood estimation of the distribution parameters. Note that the domain of a grain is defined as in the initial configuration. A total of 32 mathematical functions are used for the fitting purpose.

Figure 3 shows exemplarily the result of the fitting procedure for grains 5 and 6. The results for all grains are shown in Figure S5, Supporting Information. To assess the best fit among all grains, we use the SSE between the actual distribution and the mathematical function as the metric. The distribution of SSE values over all grains, for the 32 mathematical functions used in the current work, is shown as a box plot in **Figure 4**.

The mathematical functions that best fit the distribution of E_{vm}^{el} must be those that have low mean (denoted as $\overline{M}$) and median (denoted as $\widetilde{M}$) values of the SSE obtained from the fit to all grains. We combine the two values as $\overline{M}\widetilde{M}/(\overline{M} + \widetilde{M})$, so as to equally weight both the mean and the median values, in order to arrive at a decision on the best fit. **Table 1** lists the top five functions that result from the decision-making process, together with their standard deviation and kurtosis. The best fit is evidently obtained with the johnsonsb function, which is the bounded S_b distribution in the Johnson[67] family of distributions. A complete list of statistical values of SSE for all functions is provided in Table A1.

4.2. Peak Values in a Multimodal Distribution via GMM

For the analysis of multimodal distributions, like that in $\Delta\psi$ and E_{vm}, we need to obtain the values of individual peaks. This is
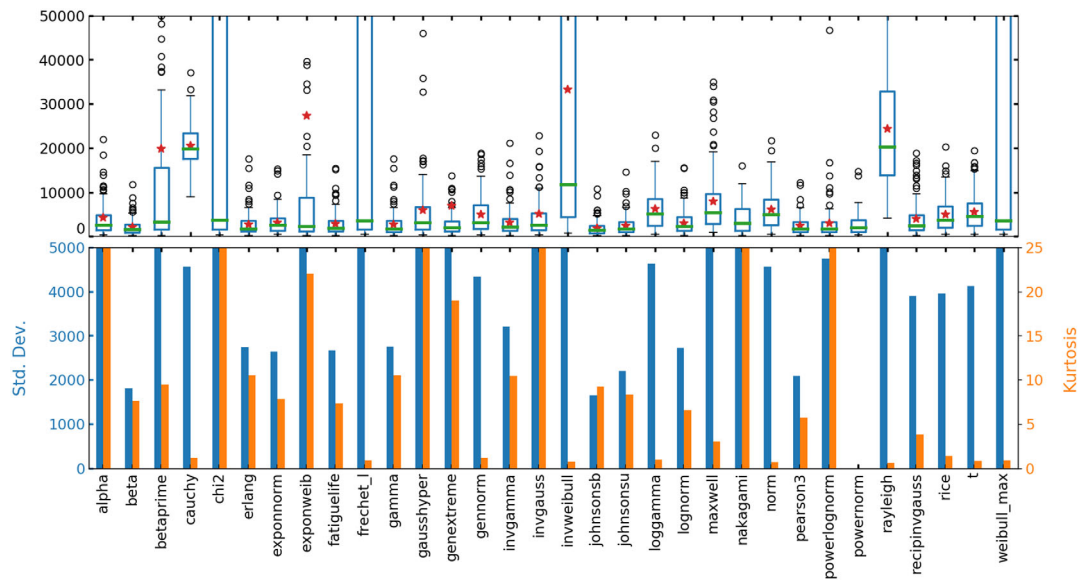
**ADVANCED
SCIENCE NEWS**
www.advancedsciencenews.com

**ADVANCED
ENGINEERING
MATERIALS**
www.aem-journal.com

Figure 4. Statistics of SSE of individual functions fit to the distributions of E_{vm}^{el} in each grain. Top row: Distribution of SSE obtained from the fit of each function over all grains as a box plot. The points marked in red denote the mean of the corresponding distribution. Bottom row: Standard deviation and kurtosis of the SSE distribution of individual functions.

Table 1. Statistics of SSE for the top five functions that result in the best fit of the distribution of E_{vm}^{el} in individual grains.

Function name	Mean [$\overline{M}$]	Median [$\widetilde{M}$]	$\frac{\overline{M}\,\widetilde{M}}{\overline{M}+\widetilde{M}}$	Std. Dev.	Kurtosis
johnsonsb	1869.76	1406.28	802.62	1653.27	9.27
beta	2103.13	1594.28	906.84	1806.70	7.61
johnsonsu	2465.08	1716.04	1011.73	2203.57	8.34
pearson3	2471.07	1786.44	1036.86	2089.86	5.73
gamma	2732.12	1780.97	1078.16	2755.70	10.49

achieved by fitting a GMM with multiple clusters—here, varying from 1 to 6—to the distributions in individual grains.

Figure 5 shows the result of such a fit, exemplarily, for two grains 29 and 118. The results for all grains are shown in Figure S6 and S7, Supporting Information. To assess the best fit, we use the Bayesian information criterion (BIC) and choose the fit that shows the least value of BIC (referred to as min (BIC)). For most grains, however, the BIC value decreases continuously with increasing number of clusters, and the best fit is obtained with six clusters (see **Figure 6**). This description, in some cases, is indicative of overfitting, or a high variance fit, as seen, for example, in grain 29 for $\Delta\psi$ (see Figure 5b).

To overcome the problem of overfitting and reduce the number of GMM clusters, we present two different approaches. For both these approaches, we first compute the clusters based on the min (BIC) criterion. We then evaluate intercluster distances in terms of the mean values of individual clusters. If this distance is less than a specific threshold, then 1) we reduce the number of clusters by one and reevaluate intercluster distances. This is done iteratively until all intercluster distances are above the specified threshold. 2) We merge all clusters whose intercluster distances

are below the threshold. The mean values, weights, and variances of each cluster are directly recalculated from the original fit.

The chosen threshold in this work is 5% of the data range of the field variable under consideration. The number of clusters recalculated by the two methods is also shown in Figure 6. It is evident that in most cases the number of clusters predicted by all three methods is the same, indicating that the optimal fit based on the least value of BIC suffices for most grains, at least in the current sample.

For cases where the min (BIC) approach results in overfitting, the two methods mentioned above result in an almost identical reduced number of clusters. Although for a few grains, the reduce method results in the fewest number of clusters, in such cases the approach seems to sometimes suffer from underfitting as seen, for example, in the distribution of $\Delta\psi$ in grain 118 (see Figure 5b).

The mean values of all clusters as predicted by the three approaches are shown in **Figure 7**. These distributions provide a general measure of the inhomogeneity of the corresponding field distribution in the polycrystalline sample. For the case of E_{vm}, two primary peaks are observed at values of 0.03 and 0.09. Further peaks are observed at approximate strain values of 0.2, 0.26, 0.36, and 0.56. For the case of $\Delta\psi$, two major peaks are observed at $\approx 3°$ and $7°$, a further significant peak is observed at $15°$, and a minor peak at roughly $25°$.

5. Discussion

The aim of the current work is to present a methodology to statistically understand and estimate distributions of continuum fields in atomistic simulations. A fundamental application of the methodology described in the current work is in obtaining an improved mechanistic understanding of the deformation behavior of polycrystalline materials. Specifically, this method
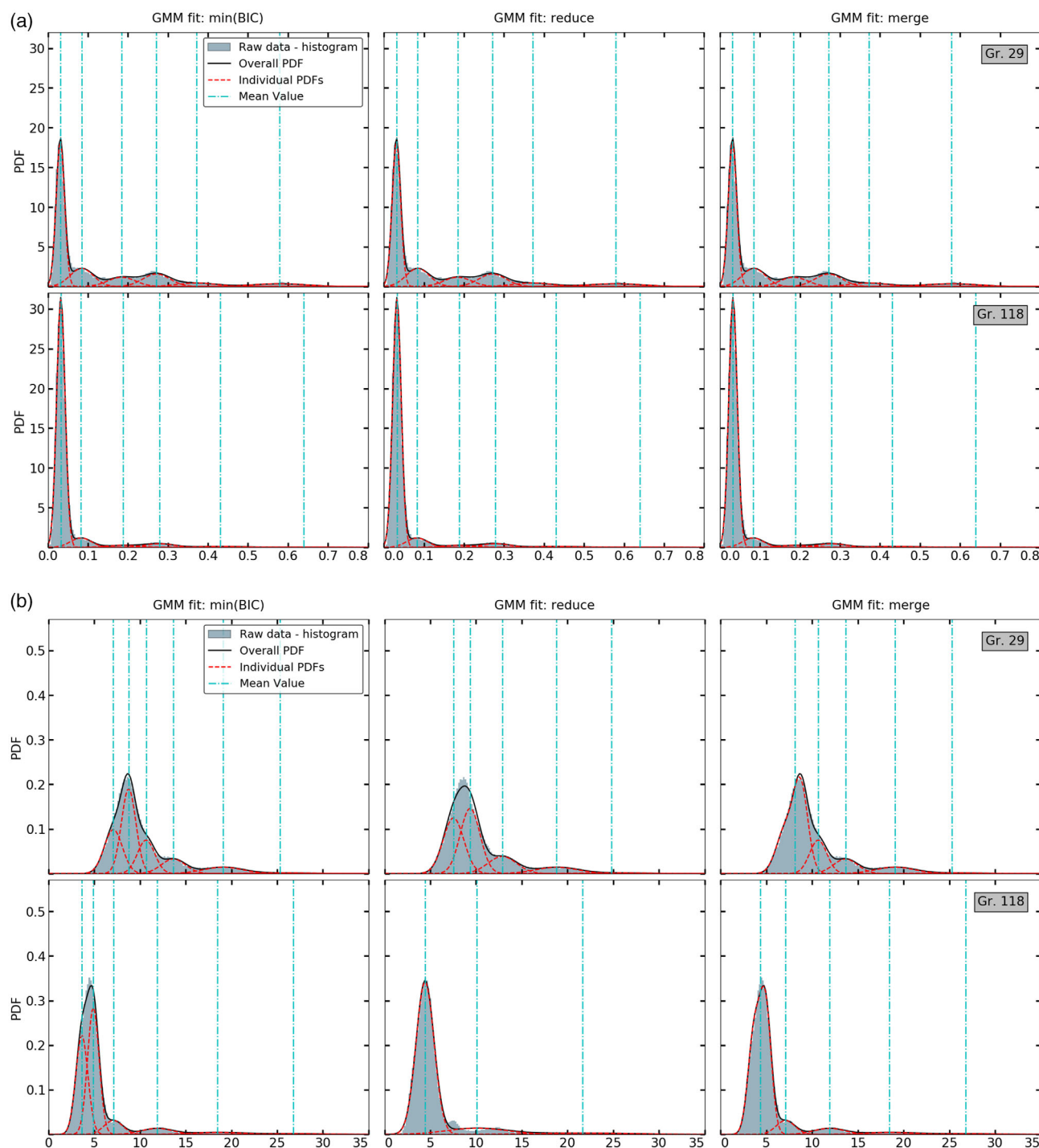
Figure 5. Results of unsupervised learning using a GMM on the multimodal distributions of E_{vm} and $\Delta\psi$. The three columns show the results of learning with min(BIC) as the evaluation criterion (left), optimum number of clusters obtained by the reduction method (center), and merge method (right).

allows us to automatically identify various deformation mechanisms like elastic strain, dislocation motion, twinning, etc. in individual grains, and quantify their relative contribution to the overall strain. Using the GMM, we have identified the individual peaks in the multimodal distributions of E_{vm} and $\Delta\psi$. These peaks essentially correspond to groups of atoms

that evidence values of E_{vm} and $\Delta\psi$ close to the mean values of the individual Gaussians. This suggests a correlation between the peaks and deformation mechanisms causing such peaks.

We illustrate this exemplarily via an a posteriori analysis of the distribution of E_{vm} in grain 65 (see **Figure 8**). Note that in the
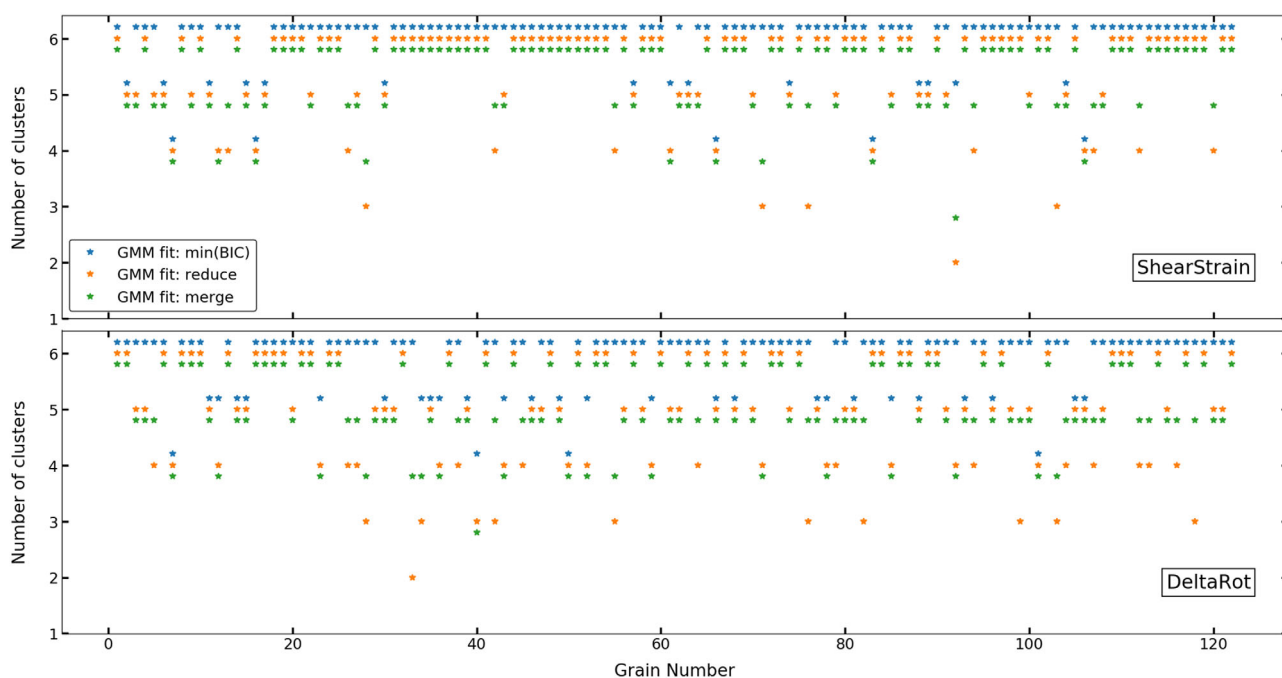
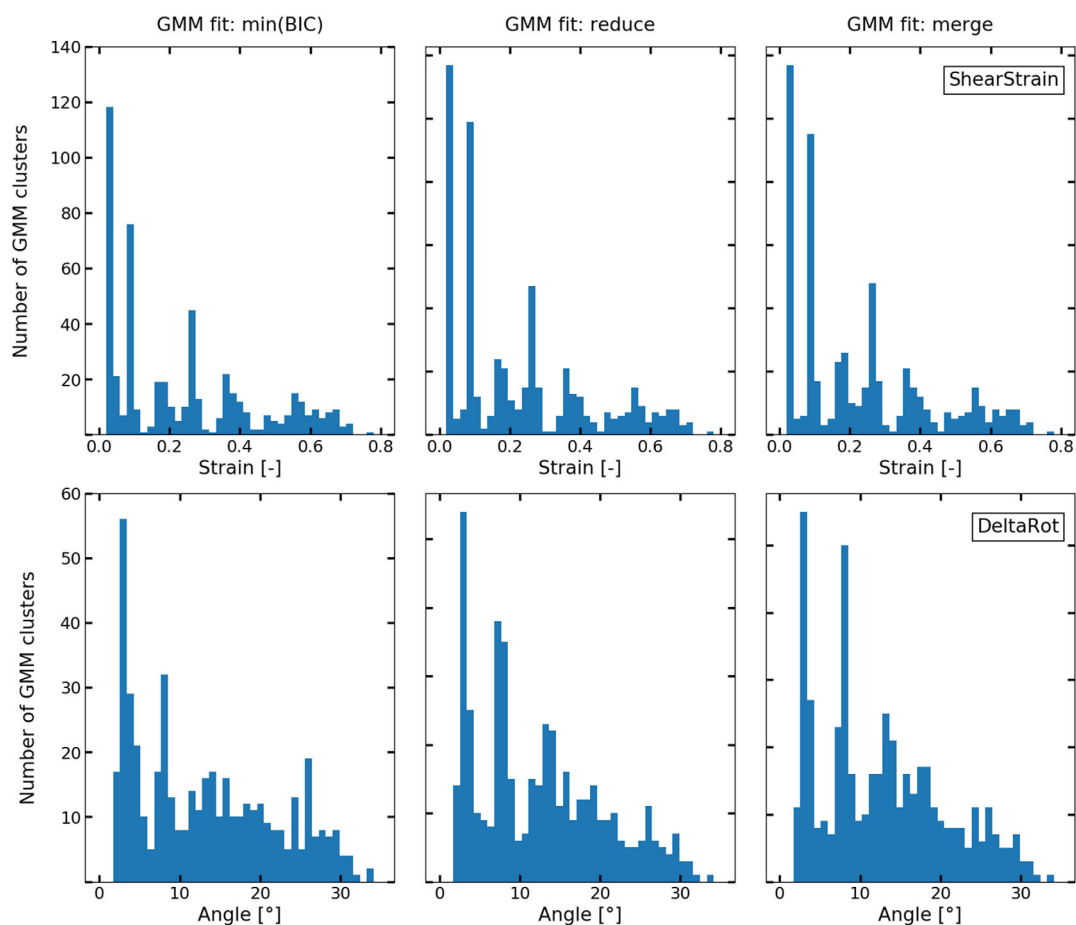**Figure 6.** Number of GMM clusters predicted in each individual grain by the three different criteria.

Figure 7. Mean values of individual GMM clusters in all grains.

**ADVANCED
SCIENCE NEWS**

www.advancedsciencenews.com

**ADVANCED
ENGINEERING
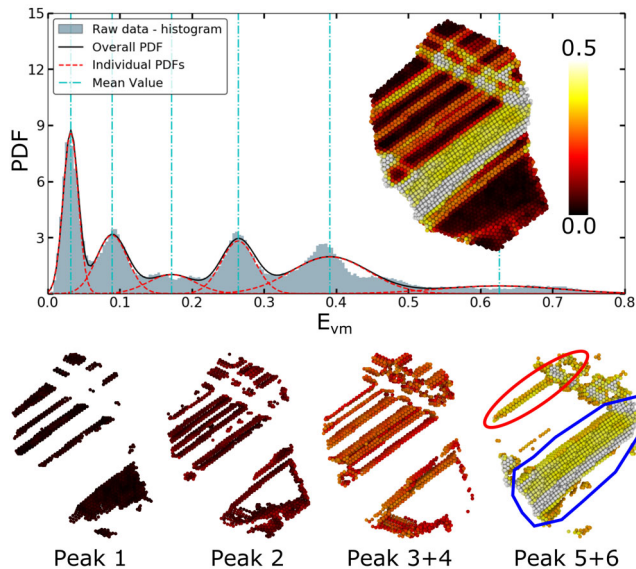MATERIALS**

www.aem-journal.com

Figure 8. Mechanisms associated with the different peaks identified by fitting GMM to the distribution of E_{vm} in grain 65. The inset in the graph shows the distribution of E_{vm} in the atomistic configuration (only a thin slice used here for the purpose of visualization). The pictures below show the atoms of the corresponding peaks.

GMM procedure, we are only interested in the values of the individual peaks and not the form of the local distribution itself, which can deviate from a Gaussian distribution. The first peak in the distribution denotes the elastic strain in the grain and is predominantly governed by atoms that have not yet experienced any dislocation activity or GB motion. The second peak is seemingly governed by atoms that form the boundary between the purely elastic region and atoms that have experienced the activity of at least one dislocation. A small contribution to this peak is also from atoms that have experienced GB migration. The third and fourth peaks denote those atoms that have seen the activity of a single dislocation. The fifth and sixth peaks denote regions with multiple dislocation activity. These include not only planes where two or more dislocations have traversed (encircled in red in Figure 8), but also twinned regions (encircled in blue in Figure 8). A small portion of these atoms, particularly those associated with peak 6, includes dislocation movement within twinned regions and the intersection nodes of slip planes.

This correlation of individual peaks in multimodal distributions, e.g., of E_{vm}, with deformation mechanisms can be generalized to the deformation behavior of nanocrystalline materials. As explained in Section 2, the multimodal nature of E_{vm} stems from the lack of dislocation generating sources inside grains. Consequently, dislocations need to be nucleated inside individual grains. Such nucleation events are usually observed at highly stressed regions close to GBs and triple junctions. Once a dislocation nucleates, it traverses through the entire grain before being absorbed at another GB. Atoms that experience the movement of such a dislocation evidence higher values of strain than atoms located far away from the dislocation activity. Merely considering the above two groups of atoms would result in two different peaks in the strain distribution: the first peak for a lower

strain value containing all atoms that have not seen any dislocation activity, and a second peak for a higher strain value containing atoms that have witnessed a dislocation activity. We hence conclude that the "fingerprints" observed in the statistics and their correlation with certain deformation mechanisms can be extended to the deformation of other nanocrystalline materials as well. The individual peaks may, however, point to different mechanisms, like, e.g., in Ag or Cu where deformation is dominated by partial dislocations.

A point must be made here on the assessment of the deformation itself, which in the current work is based on the deformation gradient (**F**)[61] as implemented in Ovito. We then use the polar decomposition $\mathbf{F} = \mathbf{RU}$ to determine the microrotation, i.e., the rotation from the initial undeformed configuration. The approach is similar to that used in the literature (see, e.g., refs. [13,36,40,68]). For the current study, this definition of the deformation gradient, **F**, is seemingly sufficient to identify the "fingerprints" in the statistics. A recent extension to calculate the higher order expansion of $\mathbf{F}$[69] may be helpful in a more detailed analysis of the deformation behavior.

A second application pertains to bridging length scales with computational frameworks using information transfer. Higher scale models such as, e.g., crystal plasticity frameworks, strain gradient models, or dislocation dynamics simulations often require information from lower length scales. These models are typically based on simplifications with respect to various aspects of the microstructure, e.g., grain boundaries, slip transfer through GBs, and influence of grain neighborhood. For instance, in classical crystal plasticity simulations, influence of GB and GB processes is usually neglected. Atomistic simulations, on the other hand, contain this information intrinsically. The distribution of field variables from atomistic simulations can be used as input in crystal plasticity frameworks to better reflect the influence of GBs and grain neighborhood on the deformation behavior in any individual grain. Such an approach is used, for instance, in the quantized crystal plasticity approach.[70] The methodology presented in the current work allows one to input distributions as mathematical functions obtained from statistics of multiple grains and configurations. For instance, the parameters of johnsonsb, i.e., the bounded S_b distribution of the Johnson family of distributions, can be used to input the distribution of elastic strain in individual grains, e.g., via FE2AT,[71] allowing for a better comparison and bridging of atomistic and continuum scale simulations. The methodology can be further extended to include, e.g., nonlocal effects and/ or to capture GB-related process explicitly by extending the data mining.

Our results show that the johnsonsb function is able to well describe the elastic strain distribution in individual grains. This function has the following mathematical form[72]

$$f(x, a, b) = \frac{b}{x(1-x)} \varphi\left(a + b \log\left(\frac{x}{1-x}\right)\right) \quad (9)$$

where the a and b are parameters with $a, b > 0$ and φ is the normal probability distribution function. It is thus evident that the distribution of elastic strain follows a lognormal nature. Taking a look at the mathematical forms of top five best fit functions tabulated in Table 1 shows in fact a close-to-lognormal

nature of the data, suggesting a multiplicative influence of sources of variation. This lognormal nature of elastic strain in nanocrystalline materials is in line with experimental observations of the total strain[73] but seemingly in contrast to the normal distribution nature of elastic strain in coarse-grained materials[74] observed after 10% applied strain in crystal plasticity simulations. This suggests a stronger coupling between elastic and plastic strains in nanocrystalline materials, and the need for more accurate constitutive models for continuum crystal plasticity simulations at the nanoscale. The coupling between elastic and plastic strains is ostensibly a result of the discrete nature of dislocation nucleation and propagation due to the absence of dislocation multiplication sources in grains.

A final point of discussion is the domain that is considered as a grain. In the current work, we estimate this domain to be composed of all atoms that formed the grain in the initial undeformed configuration. At the level of strain considered, i.e., 10% tensile strain, we do not observe any large-scale changes in the grain boundary topology. This observation is also in line with previous studies on the deformation behavior of nanocrystalline materials.[13,70,75] For specimens where large-scale GB motion is observed, and/or for larger strain values, careful consideration must be given to the domain that defines a grain. Identifying grains in deformed configurations of atomistic simulations using orientations of atoms as a measure is a challenging task.[50] Once such a domain is identified, the methodology presented in this work can be applied directly to assess the partitioning of fields in individual grains.

We hence conclude that the methodology presented in the current work can lead to improved mechanistic understanding of elastoplastic deformation behavior, and help bridge length scales in computational frameworks via information passing. The method also helps automatically identify the deformation mechanisms in individual grains and quantify the relative strain contribution of individual mechanisms. The methodology presented here can also be extended to include nonlocal effects and/or GB-related processes explicitly by improving the data mining framework.

6. Conclusions

In this work, we present a methodology for automated analysis of field distributions using statistical machine learning and data mining algorithms. The application of the algorithms is demonstrated on the distributions of elastic strain, total strain, and microrotation that have developed in individual grains in a nanocrystalline Al sample after 10% tensile strain. The distribution of elastic strain is of lognormal nature and is identified as the bounded S_b system of the Johnson family of distributions. The peak values in the multimodal distributions of total strain and microroration are identified via a GMM. We evaluate the distribution of such peak values over the entire sample and discuss the mechanistic underpinnings in terms of dislocation activity and twinning that result in such local peak values. With this method, the analysis of multiple snapshots across multiple simulations of different materials can be easily performed in an automated manner.

Appendix

Table A1. Statistics of SSE obtained for the fit of individual functions.

Function name	Mean [$\overline{M}$]	Median [$\widetilde{M}$]	$\frac{\overline{M}\,\widetilde{M}}{\overline{M}+\widetilde{M}}$	Std. Dev.	Kurtosis
alpha	4243.73	2588.33	1607.74	6321.11	54.38
beta	2103.13	1594.28	906.84	1806.70	7.61
betaprime	19799.69	3279.92	2813.80	37975.08	9.41
Cauchy	20535.29	19873.80	10099.56	4565.77	1.18
chi2	82891.06	3725.39	3565.16	342916.10	106.58
erlang	2736.39	1790.18	1082.20	2740.05	10.51
exponnorm	3149.72	2523.49	1401.02	2634.91	7.85
exponweib	27300.25	2325.51	2142.97	86182.90	22.05
fatiguelife	2878.82	1867.99	1132.89	2673.36	7.29
frechet_l	54099.54	3525.85	3310.12	86510.67	0.88
gamma	2732.12	1780.97	1078.16	2755.70	10.49
gausshyper	6029.60	3160.76	2073.71	10937.98	49.56
genextreme	6963.81	2087.40	1606.00	19587.21	18.97
gennorm	5038.20	3167.90	1944.96	4335.82	1.20
invgamma	3160.37	2131.07	1272.81	3204.16	10.43
invgauss	5134.56	2639.74	1743.42	10354.83	48.58
invweibull	33330.54	11788.77	8708.60	39006.47	0.74
johnsonsb	1869.76	1406.28	802.62	1653.27	9.27
johnsonsu	2465.08	1716.04	1011.73	2203.57	8.34
loggamma	6232.27	5063.40	2793.68	4629.79	0.97
lognorm	3060.14	2267.71	1302.50	2724.10	6.57
Maxwell	7997.27	5455.23	3243.04	7386.89	2.98
nakagami	110203.01	3024.75	2943.95	600178.01	99.75
norm	6172.50	4985.81	2758.03	4556.46	0.66
pearson3	2471.07	1786.44	1036.86	2089.86	5.73
powerlognorm	2955.93	1743.78	1096.77	4751.09	60.77
powernorm	inf	2029.85	–	–	–
Rayleigh	24373.71	20302.90	11076.42	13987.62	0.57
recipinvgauss	3936.44	2509.62	1532.56	3892.25	3.78
rice	5002.95	3680.83	2120.62	3948.73	1.37
t	5561.50	4553.68	2503.69	4120.68	0.80
weibull_max	54099.54	3525.85	3310.12	86510.67	0.88

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Keywords

[1] A. J. Hey, S. Tansley, K. M. Tolle, in *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft Research, Redmond, WA **2009**.

[2] B. G. Sumpter, R. K. Vasudevan, T. Potok, S. V. Kalinin, *NPJ Comput. Mater.* **2015**, *1*, article no. 15008.

[3] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, C. Kim, *NPJ Comput. Mater.* **2017**, *3*, article no. 54.

[4] A. Prakash, S. Sandfeld, *Pract. Metall.* **2018**, *55*, 493.

[5] S. R. Kalidindi, M. De Graef, *Annu. Rev. Mater. Res.* **2015**, *45*, 171.

[6] D. Farkas, *Curr. Opin. Solid State Mater. Sci.* **2013**, *17*, 284.

[7] M. Chassagne, M. Legros, D. Rodney, *Acta Mater.* **2011**, *59*, 1456.

[8] Z. Pan, T. J. Rupert, *Comput. Mater. Sci.* **2014**, *93*, 206.

[9] A. Vaid, J. Guénolé, A. Prakash, S. Korte-Kerzel, E. Bitzek, *Mater.* **2019**, *7*, 100355.

[10] S. Lee, A. Vaid, J. Im, B. Kim, A. Prakash, J. Guénolé, D. Kiener, E. Bitzek, S. H. Oh, *Nat. Commun.* **2020**, *11*, article no. 2367.

[11] L. A. Zepeda-Ruiz, A. Stukowski, T. Oppelstrup, V. V. Bulatov, *Nature* **2017**, *550*, 492.

[12] S. Brandstetter, H. V. Swygenhoven, S. V. Petegem, B. Schmitt, R. Maass, P. Derlet, *Adv. Mater.* **2006**, *18*, 1545.

[13] A. Prakash, D. Weygand, E. Bitzek, *Int. J. Plast.* **2017**, *97*, 107.

[14] A. Cao, Y. Wei, *Phys. Rev. B* **2007**, *76*, 024113.

[15] E. Bitzek, J. R. Kermode, P. Gumbsch, *Int. J. Fract.* **2015**, *191*, 13.

[16] D. S. Aidhy, C. Lu, K. Jin, H. Bei, Y. Zhang, L. Wang, W. J. Weber, *Acta Mater.* **2015**, *99*, 69.

[17] E. Holmström, L. Toikka, A. Krasheninnikov, K. Nordlund, *Phys. Rev. B* **2010**, *82*, 045420.

[18] J. Guénolé, A. Prakash, E. Bitzek, *Mater. Des.* **2016**, *111*, 405.

[19] J. Guénolé, A. Prakash, E. Bitzek, *Appl. Surf. Sci.* **2017**, *416*, 86.

[20] S. Kumar, X. Li, A. Haque, H. Gao, *Nano Lett.* **2011**, *11*, 2510.

[21] X. Li, X. Zhang, H. Gao, *Handbook of Materials Modeling: Applications: Current and Emerging Materials*, Springer, Cham, **2020**, p. 1845.

[22] K. Shreiber, D. Mordehai, *Modell. Simul. Mater. Sci. Eng.* **2015**, *23*, 085004.

[23] A. Sharma, J. Hickman, N. Gazit, E. Rabkin, Y. Mishin, *Nat. Commun.* **2018**, *9*, article no. 4102.

[24] F. Houllé, F. Walsh, A. Prakash, E. Bitzek, *Metall. Mater. Trans. A* **2018**, *49*, 4158.

[25] C. R. Weinberger, A. T. Jennings, K. Kang, J. R. Greer, *J. Mech. Phys. Solids* **2012**, *60*, 84.

[26] Z. Xie, J. Shin, J. Renner, A. Prakash, D. S. Gianola, E. Bitzek, *Acta Mater.* **2020**, *187*, 166.

[27] M. P. Moody, A. V. Ceguerra, A. J. Breen, X. Y. Cui, B. Gault, L. T. Stephenson, R. K. Marceau, R. C. Powles, S. P. Ringer, *Nat. Commun.* **2014**, *5*, article no. 5501.

[28] A. Prakash, J. Guénolé, J. Wang, J. Müller, E. Spiecker, M. J. Mills, I. Povstugar, P. Choi, D. Raabe, E. Bitzek, *Acta Mater.* **2015**, *92*, 33.

[29] A. Prakash, E. Bitzek, *Materials* **2017**, *10*, 88.

[30] E. N. Hahn, M. A. Meyers, *Mater. Sci. Eng. A* **2015**, *646*, 101.

[31] J. F. Panzarino, Z. Pan, T. J. Rupert, *Acta Mater.* **2016**, *120*, 1.

[32] J. Schiøtz, *Scr. Mater.* **2004**, *51*, 837.

[33] H. Van Swygenhoven, P. M. Derlet, A. G. Frøseth, *Acta Mater.* **2006**, *54*, 1975.

[34] H. Van Swygenhoven, P. Derlet, *Phys. Rev. B* **2001**, *64*, 224105.

[35] M. Dewald, W. Curtin, *Modell. Simul. Mater. Sci. Eng.* **2006**, *15*, S193.

[36] G. J. Tucker, J. A. Zimmerman, D. L. McDowell, *Int. J. Eng. Sci.* **2011**, *49*, 1424.

[37] N. Gunkelmann, I. A. Alhafez, D. Steinberger, H. M. Urbassek, S. Sandfeld, *Comput. Mater. Sci.* **2017**, *135*, 181.

[38] R. Kositski, D. Steinberger, S. Sandfeld, D. Mordehai, *Comput. Mater. Sci.* **2018**, *149*, 125.

[39] E. Bitzek, P. M. Derlet, P. M. Anderson, H. Van Swygenhoven, *Acta Mater.* **2008**, *56*, 4845.

[40] A. Gupta, J. Gruber, S. S. Rajaram, G. B. Thompson, D. L. McDowell, G. J. Tucker, *Materials* **2020**, *6*, article no. 153.

[41] G. Roberts, S. Y. Haile, R. Sainju, D. J. Edwards, B. Hutchinson, Y. Zhu, *Sci. Rep.* **2019**, *9*, 12744.

[42] W. Ma, E. J. Kautz, A. Baskaran, A. Chowdhury, V. Joshi, B. Yener, D. J. Lewis, *J. Appl. Phys.* **2020**, *128*, 134901.

[43] P. Trampert, D. Rubinstein, F. Boughorbel, C. Schlinkmann, M. Luschkova, P. Slusallek, T. Dahmen, S. Sandfeld, *Crystals* **2021**, *11*, 258.

[44] S. R. Kalidindi, *J. Appl. Phys.* **2020**, *128*, 041103.

[45] D. Steinberger, H. Song, S. Sandfeld, *Front. Mater.* **2019**, *6*, 141.

[46] B. Zheng, J. Yang, B. Liang, *J. Appl. Phys.* **2020**, *128*, 134902.

[47] M. Fernández, S. Rezaei, J. Rezaei Mianroodi, F. Fritzen, S. Reese, *Adv. Model. Simul. Eng. Sci.* **2020**, *7*, article no. 1.

[48] Y.-C. Hsu, C.-H. Yu, M. J. Buehler, *Matter* **2020**, *3*, 197.

[49] J. A. Gomberg, A. J. Medford, S. R. Kalidindi, *Acta Mater.* **2017**, *133*, 100.

[50] M. Vimal, S. Sandfeld, A. Prakash, *Materialia* **2022**, *21*, 101314.

[51] M. Ceriotti, *J. Chem. Phys.* **2019**, *150*, 150901.

[52] F. E. Bock, R. C. Aydin, C. J. Cyron, N. Huber, S. R. Kalidindi, B. Klusemann, *Front. Mater.* **2019**, *6*, 110.

[53] P.-N. Tan, M. Steinbach, V. Kumar, in *Introduction to Data Mining*, Pearson Education India, Noida, India **2016**.

[54] T. Xu, M. Li, *Philos. Mag.* **2009**, *89*, 349.

[55] P. Serrao, S. Sandfeld, A. Prakash, Optimic: A Tool to Generate Optimized Polycrystalline Microstructures for Materials Simulations, SoftwareX 15.

[56] A. Prakash, M. Hummel, S. Schmauder, E. Bitzek, *MethodsX* **2016**, *3*, 219.

[57] Y. Mishin, D. Farkas, M. J. Mehl, D. A. Papaconstantopoulos, *Mater. Res. Symp. Proc.* **1999**, *538*, 535.

[58] J. Roth, E. Eisfeld, D. Klein, S. Hocker, H. Lipp, H.-R. Trebin, *Eur. Phys. J. Special Topics* **2019**, *227*, 1831.

[59] A. Stukowski, *Modell. Simul. Mater. Sci. Eng.* **2009**, *18*, 015012.

[60] J. D. Honeycutt, H. C. Andersen, *J. Chem. Phys.* **1987**, *91*, 4950.

**ADVANCED
SCIENCE NEWS**

www.advancedsciencenews.com

**ADVANCED
ENGINEERING
MATERIALS**

www.aem-journal.com

[61] F. Shimizu, S. Ogata, J. Li, *Mater. Trans.* **2007**, *48*, 2923.

[62] Elastic Strain Calculation Modifier, OVITO 3.7.7 Manual, https://www.ovito.org/docs/current/reference/pipelines/modifiers/elastic_strain.html, (Accessed: July 2018).

[63] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, et al., *Nat. Methods* **2020**, *17*, 261.

[64] C. M. Bishop, in *Pattern Recognition and Machine Learning*, Springer, New York **2006**.

[65] W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery, *Numerical Recipes: The Art of Scientific Computing*, Cambridge University Press, Cambridge UK **2007**.

[66] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, *J. Mach. Learn. Res.* **2011**, *12*, 2825.

[67] N. L. Johnson, *Biometrika* **1949**, *36*, 149.

[68] G. J. Tucker, S. Tiwari, J. A. Zimmerman, D. L. McDowell, *J. Mech. Phys. Solids* **2012**, *60*, 471.

[69] M. S. Khorrami, J. R. Mianroodi, B. Svendsen, *J. Mech. Phys. Solids* **2022**, *166*, 104953.

[70] L. Li, P. M. Anderson, M.-G. Lee, E. Bitzek, P. Derlet, H. Van Swygenhoven, *Acta Mater.* **2009**, *57*, 812.

[71] J. J. Möller, A. Prakash, E. Bitzek, *Modell. Simul. Mater. Sci. Eng.* **2013**, *21*, 055011.

[72] Scipy Stats Package, https://docs.scipy.org/doc/scipy/reference/stats.html, (Accessed: March 2021).

[73] A. Tang, H. Liu, G. Liu, Y. Zhong, L. Wang, Q. Lu, J. Wang, Y. Shen, *Phys. Rev. Lett.* **2020**, *124*, 155501.

[74] J. Chen, A. M. Korsunsky, *Mater. Des.* **2020**, *198*, 109319.

[75] E. Bitzek, C. Brandl, P. Derlet, H. Van Swygenhoven, *Phys. Rev. Lett.* **2008**, *100*, 235501.

2200574 (13 of 13)