PROCEEDINGS A

royalsocietypublishing.org/journal/rspa

Research

Cite this article: Zhang Y, Dunham SD, Bronkhorst CA, Chen N. 2025 Physics-assisted data-driven stochastic reduced-order models for attribution of heterogeneous stress distributions in low-grain polycrystals. *Proc. R. Soc. A* **481**: 20240898. https://doi.org/10.1098/rspa.2024.0898

Received: 20 November 2024
Accepted: 7 February 2025

Subject Areas:
applied mathematics, mathematical modelling, materials science

Keywords:
causal inference, void nucleation, statistical modelling

Author for correspondence:
Nan Chen
e-mail: chennan@math.wisc.edu

Physics-assisted data-driven stochastic reduced-order models for attribution of heterogeneous stress distributions in low-grain polycrystals

Yinling Zhang[1], Samuel D. Dunham[2], Curt A. Bronkhorst[2] and Nan Chen[1]

[1]Department of Mathematics and [2]Department of Mechanical Engineering, University of Wisconsin, Madison, WI 53706, USA

YZ, 0000-0001-8636-3483

Understanding stress distributions at grain boundaries in polycrystalline materials is crucial for predicting damaged nucleation sites. In high-purity materials, voids often nucleate at grain boundaries due to high stress from granular interactions and weakened atomic ordering. While traditional crystal plasticity models simulate grain-level mechanics, their high computational cost often prevents systematic identification of critical microstructural features and efficient forecast of extreme damage events. This paper addresses these challenges by developing a computationally efficient physics-assisted statistical modelling framework. The method starts by leveraging physical knowledge to hypothesize a broad set of microstructural factors influencing stress conditions. Causal inference is then applied to reveal the predominant features with physical explanations, leading to a parsimonious statistical model. A conditional Gaussian mixture model (CGMM) is employed when the identified relationship is utilized as a predictive model to quantify the uncertainty not readily explained by these features. Using body-centred cubic (BCC) tantalum as a representative material, a series of synthetic microstructures from single- to octu-crystal configurations are created. Results show that high-stress states strongly correlate with the elastic and plastic deformation capabilities

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.7679513.

THE ROYAL SOCIETY
PUBLISHING

and the directional misalignment of grain responses near boundaries. The statistical model achieves rapid and accurate forecasts, demonstrating its potential for analysing realistic polycrystalline materials.

1. Introduction

The behaviour of polycrystalline metallic materials under extreme loading conditions is critical in material science and engineering. When the local stress state exceeds the local void nucleation strength, a void will nucleate, which can progress to catastrophic material failure through void growth and coalescence processes [1–6]. These voids are discrete localized features that first form at tiny length scales before contributing to a larger field of voids with larger length scales. The formation of each void is considered an extreme event in response to the loading of an aggregate composite polycrystalline metal. For high-purity polycrystalline materials, experimental evidence has consistently shown that these voids preferentially nucleate at grain boundaries [3,6–9]. This preferential nucleation at grain boundaries stems from elevated stress states due to incompatible deformation between neighbouring grains [10–14]. Stress concentration is further exacerbated by the inherently weak atomic structure in these boundary regions, making them particularly susceptible to void formation [15–17]. Given the critical role of grain-boundary stress states in void nucleation, modelling and quantifying localized concentrations for reliable prediction and prevention of damage in polycrystalline materials become crucial.

Current models for representing void nucleation and growth under dynamic loading conditions can be broadly summarized into several categories. Early work by Johnson [18] introduced a mathematical model for the growth of voids under tensile mean stress and applied it to spallation problems through a microscopic to continuous framework. This model was further advanced by incorporating micro-inertial effects in several works [19–21]. More recent developments have demonstrated the probability distribution of void nucleation [5,22]. The work of [7] utilized a soft-coupled linkage technique between the macroscale damage model and micromechanical calculations to study the nucleation of voids. Note that maximum stress intensity usually occurs at the grain boundary [3,6] because these regions already exhibit inherently weakened atomic structures [23,24], making them most susceptible to damage initiation. These extreme events not only drive damage initiation and evolution in materials but also play crucial roles across many different types of physical systems where accurate prediction of extreme scenarios is essential for reliability and safety considerations [25,26]. However, the complicated nonlinear mechanisms and the intrinsic heterogeneity of stress states near grain boundaries pose unique challenges in predicting void nucleation, where stress distributions in these regions often exhibit complex non-Gaussian characteristics [15,16,27]. The tail of the non-Gaussian distributions often corresponds to extreme damage events, which are difficult to capture using conventional mean-field approaches. Furthermore, the high computational cost of traditional models does not allow us to carry out a large number of simulations, which prevents a quantitative understanding of the processes and the efficient forecasting of extreme damage events. Developing efficient methods to explicitly identify the critical variables contributing to elevated stress states not only facilitates the physical understanding of these mechanisms but also advances rapid and accurate forecasting of extreme stress values near grain boundaries in polycrystalline materials.

In this paper, a physics-assisted statistical reduced-order modelling approach is developed to identify and quantify the key factors that control stress concentration at grain boundaries, which is a critical precursor to damage initiation. Rather than attempting to resolve all microscopic mechanisms, which is computationally prohibitive for engineering applications, this approach establishes quantitative relationships between elevated stress states and a small set of dominant, easily estimable microstructural features. Using body-centred cubic (BCC) tantalum as a model system for simulation, the approach is demonstrated through analysis of increasingly complex synthetic microstructures, from single-crystal to octu-crystal configurations. The method begins by leveraging physical insights to hypothesize a comprehensive set of microstructural factors influencing stress conditions. Causation entropy [28–32] is then employed to identify a parsimonious relationship between stress values and a select few dominant features with clear physical interpretations. Finally, a conditional Gaussian mixture model (CGMM) [33] is developed to quantify the uncertainty that is not explained by the selected primary features.

The physics-assisted statistical modelling approach has several desirable features in analysing complex microstructural systems. First, traditional methods often struggle with a vast number of factors influencing stress states, which rely on either physical intuition alone or purely data-driven approaches. By contrast, the proposed method combines physical knowledge with statistical methods grounded in information theory. Second, by identifying causal relationships among numerous microstructural features suggested by physical intuition, it pinpoints a small set of dominant mechanisms governing stress concentration from the high-dimensional feature space. Compared to a full-field crystal plasticity calculation from a comprehensive computational model, obtaining the values of these features is of much lower cost. Therefore, the approach reduces simulation costs by orders of magnitude while preserving the essential physics and non-Gaussian statistics of grain-boundary interactions. Third, the CGMM provides appropriate uncertainty quantification for the identified model and its prediction. Such statistical characterization is essential for predicting extreme events, such as damage occurrence, in a robust probabilistic way. The proposed approach is computationally efficient and facilitates physical interpretability. It advances the understanding of microstructures with enhanced damage resistance, which extends beyond stress prediction to inform material design strategies.

The remainder of the paper is organized as follows. In §2 we present the general data-driven statistical reduced-order modelling framework. In §3 we introduce the polycrystal model and definitions of microstructural features in this study. In §4 we introduce simulation configurations and corresponding settings. In §5 we identify critical factors governing elevated stress states and quantifying prediction uncertainties. Finally, in §6, we present the discussion and conclusion.

2. Physics-assisted data-driven statistical reduced-order modelling framework

The statistical reduced-order modelling framework developed here aims to identify the primary variables that lead to stress conditions conducive to void nucleation. The approach combines statistical tests, information-theoretic measures and advanced probabilistic modelling techniques to reveal the underlying mechanism contributing to the complex stress states near grain boundaries. Except the single-crystal case, which focuses on the resultant stress, the maximum stress in the field from each simulation will be utilized as the target variable in all the other tests. Developing the statistical model requires a set of crystal plasticity simulations from traditional computational models. Nevertheless, once the statistical model is developed, the forecast, which relies only on computing the few selected features without running the original computational models, becomes highly efficient. The framework consists of two phases: data processing and model identification with feature analysis.

The data processing phase involves:

(i) Compute crystal plasticity simulations under different initial loading conditions for increasingly complex idealized microstructure configurations.

(ii) Apply the Kolmogorov–Smirnov (KS) test to ensure the statistical significance of sampling simulations in each configuration.

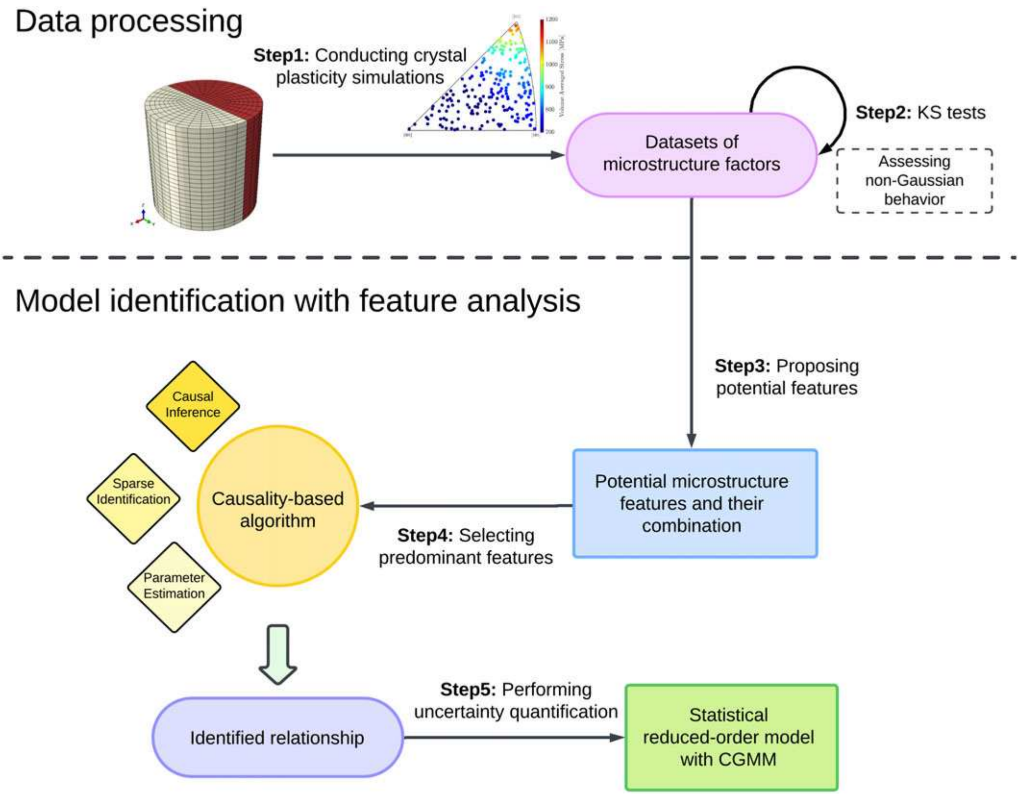
Data processing
Step1: Conducting crystal plasticity simulations
Datasets of microstructure factors
Step2: KS tests
Assessing non-Gaussian behavior
Model identification with feature analysis
Step3: Proposing potential features
Potential microstructure features and their combination
Step4: Selecting predominant features
Causality-based algorithm
Causal Inference
Sparse Identification
Parameter Estimation
Identified relationship
Step5: Performing uncertainty quantification
Statistical reduced-order model with CGMM

Figure 1. Overview of the physics-assisted data-driven stochastic reduced-order modelling framework.

The model identification with feature analysis phase consists of:

(iii) Construct a comprehensive candidate library of potential microstructural features based on the physical understanding of deformation mechanisms and grain-boundary interactions.

(iv) Apply causation entropy analysis to systematically identify the principal factors contributing to stress states (e.g. the maximum stress), leading to a physically interpretable model.

(v) Characterize model uncertainties using CGMMs to advance the statistical forecast of extreme events.

Figure 1 includes a schematic diagram illustrating the analytical framework of our study. It outlines the key steps from initial polycrystal simulations through statistical analysis to the identification of primary microstructural factors influencing stress states. The following sections detail each component of the statistical and computational methodology, including the KS tests, causation entropy analysis and the application of CGMMs for uncertainty quantification.

(a) KS test

The analysis of stress distribution in polycrystalline materials requires enough samples of simulated configurations. As mentioned above, the non-Gaussian distribution of maximum stress resulting from various initial loading conditions is crucial for damage prediction. Adopting an adequate sampling size is the prerequisite for unbiasedly characterizing such non-Gaussian statistics. To this end, the so-called KS test is utilized [34–36] to examine the statistics of the data and determine a minimum number of samples that leads to statistically significant results.

The KS test is a non-parametric test that can directly evaluate whether the data follows a specific continuous probability distribution. For the specific application here, the KS test facilitates determining whether the distribution of the sample data is statistically significantly different from a normal distribution.

For each idealized configuration, multiple simulations with varying crystallographic orientations are performed. The maximum stress data from these simulations are normalized to enable direct comparison with the normal distribution. By progressively increasing the number of simulations, the KS test examines whether the distribution of the stress values is statistically distinguishable from the standard normal distribution. The test statistic D quantifies the maximum difference between the empirical distribution of normalized stress states and the standard normal distribution

$$D = \sup_x |F_n(x) - F(x)|, \tag{2.1}$$

where $F_n(x)$ is the empirical distribution function of the sample and $F(x)$ is the cumulative distribution function of the reference distribution [37], which is a normal distribution in this study.

For each sample size, the test statistic D is compared with a critical value that depends on the significance level and sample size. If D exceeds the critical value which is used for the chosen significant level, the hypothesis of normality is rejected. This indicates the significant difference between the distribution of stress states and the normal distribution. As more simulations are included in these KS tests, the reject rate will systematically increase since the non-Gaussian feature will be clearer as more samples are included. If the rejection rate exceeds 95%, then it implies that the sampling size is sufficient to consistently detect non-Gaussian feature of the stress states. At this point, the sample size is considered adequate for capturing the statistical properties of the stress distribution, enabling reliable subsequent analyses.

The rigorous assessment of sampling adequacy by the KS test is important for subsequent analysis of microstructural effects on stress states. Since sufficient sampling captures statistical characteristics, it is a critical prerequisite for reliable damage prediction analysis. Specifically, the causation entropy method employed in the next subsection requires adequate data to accurately identify causal relationship between the multiple microstructural factors and stress states. Insufficient sampling may lead to wrong or missing causal connections, affecting the identification of key physical mechanisms governing stress concentration at grain boundaries.

(b) Causality-based learning algorithm

The complex relationship between microstructural features and stress states presents significant challenges for traditional analysis methods. While numerous microstructural characteristics potentially influence stress concentration at grain boundaries, not all correlations indicate causal relationships. For example, two features might show a strong correlation simply because they are both affected by a common underlying mechanism rather than having a direct causal relationship [38,39]. Furthermore, the nonlinear interactions between multiple features make it challenging to identify genuinely influential factors through conventional correlation analysis. Therefore, a causality-based learning algorithm is developed to systematically identify microstructural features contributing information to stress states to overcome these challenges. There are several advantages over traditional methods based on information theory. First, the algorithm distinguishes between direct causal relationships and indirect correlations. Thus, it identifies the fundamental mechanisms driving stress concentration. Second, it naturally accounts for complex interactions between multiple features, providing a complete understanding of the physical system. Third, it allows for efficient processing of a large number of potential factors while maintaining physical interpretability. The implementation of this framework consists of three main components: construction of a physics-based feature library, identification of causal relationships through causation entropy, and parameter estimation of the resulting identified relationship between the stress values of the selected features.

(i) Candidate feature library

A library **f** containing M possible candidate factors is constructed to model the relationship between microstructural factors and stress states

$$\mathbf{f}=\{f_1,\ldots,f_{m-1},f_m,f_{m+1},\ldots,f_M\}. \tag{2.2}$$

This library is developed based on physical knowledge and expert insights into the behaviour of polycrystalline material. It includes a wide range of functions to cover various potential relationships between microstructural features (such as grain orientation, elastic constants and non-Schmid factors) and stress states. This also allows for proposing new potential physical relationships that are not obvious.

(ii) Computing the causation entropy

The foundation of our analysis lies in the information theory. For multi-dimensional variables **X** and **Y**, the fundamental entropy, conditional entropy and joint entropy are defined as

$$\left.\begin{aligned} H(\mathbf{X}) &= -\int_{\mathbf{x}} p(\mathbf{x})\log(p(\mathbf{x}))\,\mathrm{d}\mathbf{x}, \\ H(\mathbf{Y}|\mathbf{X}) &= -\int_{\mathbf{x}}\int_{\mathbf{y}} p(\mathbf{x},\mathbf{y})\log(p(\mathbf{y}|\mathbf{x}))\,\mathrm{d}\mathbf{y}\,\mathrm{d}\mathbf{x} \\ \text{and}\qquad H(\mathbf{X},\mathbf{Y}) &= -\int_{\mathbf{x}}\int_{\mathbf{y}} p(\mathbf{x},\mathbf{y})\log(p(\mathbf{x},\mathbf{y}))\,\mathrm{d}\mathbf{y}\,\mathrm{d}\mathbf{x}. \end{aligned}\right\} \tag{2.3}$$

Building upon these foundational measures, the causation entropy $C_{f_m\to\sigma_n|[\mathbf{f}\backslash f_m]}$ is introduced to evaluate how each microstructural feature f_m influences the stress state σ_n in our nth simulation [28–32]

$$C_{f_m\to\sigma_n|[\mathbf{f}\backslash fm]}=H(\sigma_n|[\mathbf{f}\backslash fm])-H(\sigma_n|\mathbf{f}). \tag{2.4}$$

Here, $\mathbf{f}\backslash f_m$ represents the set of all candidate features excluding f_m, and $H(\cdot|\cdot)$ denotes the conditional entropy. The causation entropy formulation in equation (2.4) quantifies the unique information that feature f_m contributes to explaining the stress state σ_n, beyond what is already captured by all other features. This approach extends beyond traditional correlation analysis in a fundamental way. While correlation merely measures the statistical relationship between two variables, causation entropy accounts for the complex interactions among all features in the candidate library. For example, if a common factor f'_m influences both σ_n and f_m, these variables may exhibit strong correlation despite lacking a direct causal relationship. In such cases, while the correlation between σ_n and f_m might be high, the causation entropy $C_{f_m\to\sigma_n|[\mathbf{f}\backslash f_m]}$ would correctly identify the absence of direct causation by yielding a value near zero.

The practical implementation of this framework faces a significant challenge: high-dimensional numerical integration required by equations (2.3). This issues could be solved through a Gaussian approximation strategy that transforms the problem into a tractable form

$$\begin{aligned} C_{\mathbf{Z}\to\mathbf{X}|\mathbf{Y}} &= H(\mathbf{X}|\mathbf{Y})-H(\mathbf{X}|\mathbf{Y},\mathbf{Z}) \\ &= H(\mathbf{X},\mathbf{Y})-H(\mathbf{Y})-H(\mathbf{X},\mathbf{Y},\mathbf{Z})+H(\mathbf{Y},\mathbf{Z}) \\ &= \frac{1}{2}\ln(\det(\mathbf{R}_{\mathbf{XY}}))-\frac{1}{2}\ln(\det(\mathbf{R}_{\mathbf{Y}}))-\frac{1}{2}\ln(\det(\mathbf{R}_{\mathbf{XYZ}})) \\ &\quad+\frac{1}{2}\ln(\det(\mathbf{R}_{\mathbf{YZ}})). \end{aligned} \tag{2.5}$$

Here, $\mathbf{R}_{\mathbf{XYZ}}$ represents the covariance matrix of variables $(\mathbf{X},\mathbf{Y},\mathbf{Z})$, with similar definitions for other covariance terms.

The Gaussian approximation expressed in equation (2.5) provides an efficient computational framework for evaluating causation entropy, particularly suitable for systems with moderately large dimensions such as the low-order system with leading moment equations. While this approximation might introduce some error when the true distribution significantly deviates from Gaussian behaviour, such precision is not critical for our primary objective. Rather than seeking exact entropy values, the goal is to detect whether the causation entropy $C_{f_m \to \sigma_n | [\mathbf{f} \backslash f_m]}$ exceeds a small threshold value, thereby identifying meaningful causal relationships. This approach has proven effective in practice, as significant causal relationships detected in higher-order moments typically manifest reliably in the Gaussian approximation. Furthermore, this method enables efficient identification of sparse model structures, where the precise coefficients can be subsequently determined through linear regression, as discussed later. The reliability of this Gaussian approximation approach is well-established with many applications [40–44].

Once the model is determined by selecting a small set of candidate functions, the coefficients can be estimated via a least-squares method.

(c) Conditional Gaussian mixture modelling

Although the identified causal relationships can capture most non-Gaussian features, they are deterministic and cannot directly quantify the uncertainty when predicting extreme events. Moreover, the difference between the identified model and true maximum stress (i.e. residual) in experiments may exhibit non-Gaussian features that prevent them from being treated as simple white noise. To address these challenges, a CGMM is introduced that builds upon both the identified causal relationships and non-Gaussian features of stress states. The general Gaussian mixture model (GMM) is a suitable approach for representing non-Gaussian features in the form of a combination of multiple Gaussian distributions [33,45]. While the GMM, describing the total residual, can be used directly as a crude quantification of the forecast uncertainty, the maximum stress values vary significantly depending on different physical conditions. Therefore, it is advantageous to develop a modified version of the GMM, namely, the CGMM, which characterizes the differences in the forecast uncertainty, conditioned on different physical conditions, allowing a more refined uncertainty quantification. The CGMM naturally accounts for the heavy tails associated with extreme events [46,47].

The CGMM approach consists of two main steps:

(i) *Training phase*: collecting pairs of data consisting of (i) the results from the identified physics-based model σ_{model} and (ii) the corresponding true maximum stress values obtained from simulations σ_{truth}. These pairs are used to construct a GMM that captures the joint distributions of the true and estimated maximum stress values. Then, conditioning on each σ_{model}, the corresponding GMM describing σ_{truth}, known as the CGMM, can be formulated.

(ii) *Prediction phase*: given a new result σ^*_{model} from the identified model, the predicted maximum stress σ^*_{truth} is obtained based on the updated parameters from the corresponding CGMM. This estimation provides not just a single predicted value but a range of possible values with their associated probabilities.

The details for these two steps are introduced in §2c(i) and §2c(ii).

(i) Training part

The training phase begins with collecting pairs of data from the identified physics-based model and the true value for maximum stress. Let σ_{model} represent the predictions from the physics-based model, and σ_{truth} denote the true maximum stress values from simulations. These pairs of data are used to construct joint probability distribution using GMM.

A GMM represents a probability distribution as a weighted sum of multiple Gaussian component distributions. Each component in the mixture is defined by its mean, covariance matrix and mixture weight, where the weights sum to unity. The joint distribution of σ_{model} and σ_{truth} is given by

$$p(\sigma_{\text{model}}, \sigma_{\text{truth}}) = \sum_{k=1}^{K} w_k \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{2.6}$$

where K is the number of Gaussian components, w_k are the mixture weights and $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ represents the bivariate Gaussian distribution with mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$.

The parameters of each Gaussian component are

$$\boldsymbol{\mu}_k = (\mu_M^k, \mu_T^k) \quad \text{and} \quad \boldsymbol{\Sigma}_k = \begin{bmatrix} \Sigma_{MM}^k & \Sigma_{MT}^k \\ \Sigma_{TM}^k & \Sigma_{TT}^k \end{bmatrix}, \tag{2.7}$$

where M and T stand for model and truth, respectively. In this study, the vector $\boldsymbol{\mu}_k$ is in two dimensions, and $\boldsymbol{\Sigma}_k$ is a 2×2 matrix.

Before estimating the mean vector and the covariance matrix, the number of Gaussian components K is a critical parameter in constructing the GMM. It is typically determined based on the characteristics of the data and the specific application requirements. Selecting an appropriate K involves balancing model complexity and accuracy while avoiding overfitting. Several approaches can be employed to determine the optimal number of components, including information criteria methods such as the Akaike information criterion, the Bayesian information criterion [48–50] and direct cross-validation [51] which is used in this study.

Once the number of components is determined, the other parameters, such as mixture weights, means and the covariance matrix, can be optimized using the expectation-maximization (EM) algorithm [52]. Since it balances computational efficiency and estimation accuracy well, it is particularly suitable for moderate-sized datasets. The EM algorithm alternates between component assignments and updating the component parameters. For each step, the log-likelihood is calculated. This iterative process continues until the log-likelihood converges, indicating that a local optimum for the model parameters is found. The resulting GMM parameters capture both the overall distribution of stress values and the complex relationships between the identified model and true values across different physical conditions.

(ii) Prediction part

We can use the model to predict new cases after obtaining the GMM parameters from the training phase. Given a new estimation σ^*_{model} from the identified physics-based model, the goal is to estimate the corresponding maximum stress σ^*_{truth}. This is achieved through the conditional distribution of the GMM, namely, the CGMM.

The distribution of σ^*_{truth} conditioned on σ^*_{model} is given by

$$p(\sigma^*_{\text{truth}} \mid \sigma^*_{\text{model}}) = \sum_{k=1}^{K} \gamma_k \mathcal{N}(\tilde{\mu}^k_{T|M}, \tilde{\Sigma}^k_{T|M}), \tag{2.8}$$

where the updated weights γ_k, means $\tilde{\mu}^k_{T|M}$ and covariances $\tilde{\Sigma}^k_{T|M}$ for each component are computed using the basic property of joint Gaussian distribution in light of equation (2.7). The γ_k is given by

$$\gamma_k = \frac{\pi_k \mathcal{N}(\sigma^*_{\text{model}} \mid \mu_M^k, \Sigma_{MM}^k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\sigma^*_{\text{model}} \mid \mu_M^j, \Sigma_{MM}^j)}, \tag{2.9}$$

where $\mathcal{N}(\sigma^*_{\text{model}} \mid \mu_M^k, \Sigma_{MM}^k)$ is the probability density of observing σ^*_{model} under the kth Gaussian component. The responsibility γ_k represents how much each Gaussian component contributes to predicting σ^*_{truth} given a new model result σ^*_{model}.

It should be noted that the responsibility γ_k calculated using equation (2.9) is not the most accurate representation of the contribution of each Gaussian component to the true stress σ^*_{truth}. This is because γ_k is determined solely based on the estimated stress of the identified model

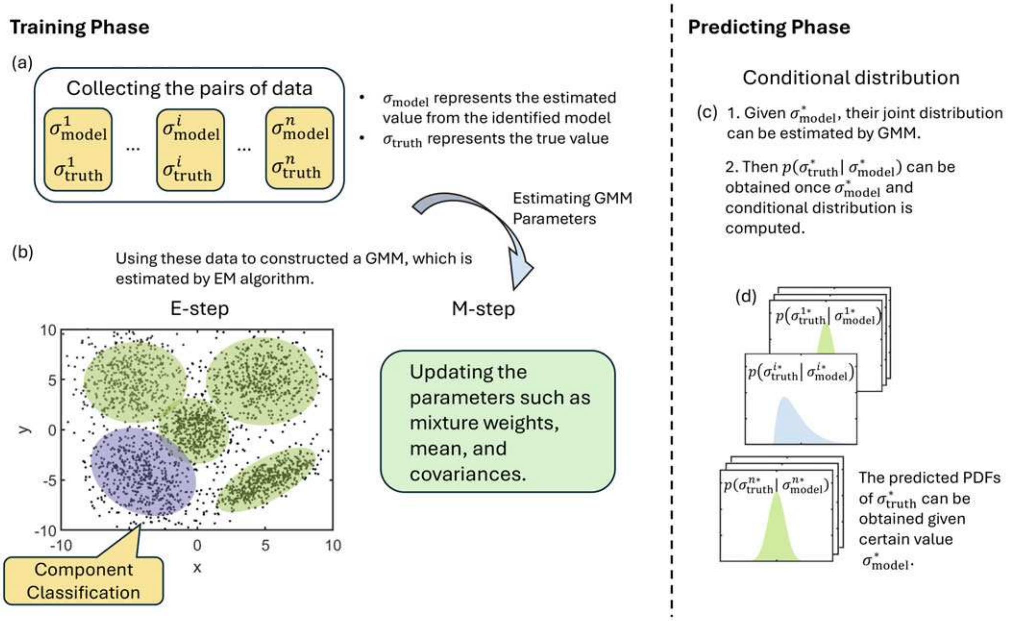
Training Phase
(a)
Collecting the pairs of data
$\sigma^1_{\mathrm{model}}$ $\sigma^1_{\mathrm{truth}}$ ··· $\sigma^i_{\mathrm{model}}$ $\sigma^i_{\mathrm{truth}}$ ··· $\sigma^n_{\mathrm{model}}$ $\sigma^n_{\mathrm{truth}}$
• σ_{model} represents the estimated value from the identified model
• σ_{truth} represents the true value
Estimating GMM Parameters
(b)
Using these data to constructed a GMM, which is estimated by EM algorithm.
E-step
M-step
Updating the parameters such as mixture weights, mean, and covariances.
10 5 0 -5 -10
y
-10 0 5 10
x
Component Classification
Predicting Phase
Conditional distribution
(c) 1. Given $\sigma^*_{\mathrm{model}}$, their joint distribution can be estimated by GMM.
2. Then $p(\sigma^*_{\mathrm{truth}} | \sigma^*_{\mathrm{model}})$ can be obtained once $\sigma^*_{\mathrm{model}}$ and conditional distribution is computed.
(d)
$p(\sigma^{1*}_{\mathrm{truth}} | \sigma^{1*}_{\mathrm{model}})$
$p(\sigma^{i*}_{\mathrm{truth}} | \sigma^{i*}_{\mathrm{model}})$
$p(\sigma^{n*}_{\mathrm{truth}} | \sigma^{n*}_{\mathrm{model}})$
The predicted PDFs of $\sigma^*_{\mathrm{truth}}$ can be obtained given certain value $\sigma^*_{\mathrm{model}}$.

Figure 2. The overview diagram of conditional Gaussian mixture modelling.

$\sigma^*_{\mathrm{model}}$ and the parameters (μ^k_M, Σ^k_{MM}) associated with σ_{model}. A more precise calculation of the responsibility would involve considering the joint probability distribution of both σ_{model} and σ_{truth}. Nevertheless, in the absence of the true stress value for a new data point, the responsibility γ_k calculated using equation (2.9) serves as a reasonable approximation of the contribution of each Gaussian component to the prediction of $\sigma^*_{\mathrm{truth}}$.

The parameters for the distribution $p(\sigma^*_{\mathrm{truth}} \,|\, \sigma^*_{\mathrm{model}})$ is given by the following based on the conditional distribution theory of a joint Gaussian distribution [53]:

$$\left.\begin{aligned} \tilde{\mu}^k_{T|M} &= \mu^k_T + \Sigma^k_{TM}(\sigma^*_{\mathrm{model}} - \mu^k_M)/\Sigma^k_{MM} \\ \text{and} \qquad \tilde{\Sigma}^k_{T|M} &= \Sigma^k_{TT} - \Sigma^k_{TM}\Sigma^k_{MT}/\Sigma^k_{MM}. \end{aligned}\right\} \tag{2.10}$$

Note that these two parameters are both scalars.

For each given $\sigma^*_{\mathrm{model}}$, a predicted probability density function (PDF) is obtained for the corresponding true stress $\sigma^*_{\mathrm{truth}}$. This predicted PDF provides valuable information about the uncertainty in estimating $\sigma^*_{\mathrm{truth}}$. Instead of a single-value estimate, the PDF gives a range of possible values for $\sigma^*_{\mathrm{truth}}$ along with their associated probabilities. It serves as a powerful tool for quantifying uncertainty in the model predictions. It also enables us to make more informed decisions and assess the reliability of the model results.

Figure 2 summarizes the process of the CGMM framework.

3. Material microstructures and computational models

(a) Computational crystal model

The crystal plasticity model used in this work is formulated within a large deformation framework and incorporates non-Schmid effects characteristic of BCC materials such as tantalum. The model is highly nonlinear and comprehensive, with sophisticated physical mechanism representation. We note that the model presented herein is a local material model that does not consider gradients of field variables in the formulation. Models that include gradient effects can more accurately account for the atomistic nature of grain boundaries, as dislocation densities on slip systems will experience a discontinuity across the boundary and will thus feel the grain boundary provided that the material point is sufficiently close [54]. Many non-local crystal plasticity models incorporate the effect of statistically stored dislocations as well as geometrically necessary dislocations [55–58]. Geometrically necessary dislocations are those that accommodate plastic strain gradients in the material, while statistically stored dislocations are responsible for the evolving structural resistance. Since we seek to construct a statistical model formulated on a statistically significant number of observations, several thousand calculations were performed as will be shown in §5. It is thus probable that the computational expense to employ a non-local material model would be impractical, thus we adopt a local material model. The main model components are summarized hereafter, and the model details can be found in [7,59,60].

(i) Kinematics

First, the deformation gradient is cast in a multiplicative manner into elastic and plastic parts,

$$\mathbf{F} = \nabla\boldsymbol{\phi} = \mathbf{F}^e\mathbf{F}^p, \tag{3.1}$$

where $\boldsymbol{\phi}$ is the motion in terms of the position in the reference configuration and here $\nabla(\bullet)$ denotes a derivative in the reference configuration. The plastic deformation gradient evolves with its velocity gradient,

$$\dot{\mathbf{F}}^p = \mathbf{L}^p\mathbf{F}^p \quad \text{and} \quad \mathbf{L}^p = \sum_{\alpha} \dot{\gamma}_p^{\alpha}\mathbf{S}_0^{\alpha}, \tag{3.2}$$

where $\mathbf{L}^p$ is the plastic velocity gradient, $\dot{\gamma}_p^{\alpha}$ is the shear rate on slip system α, $\mathbf{S}_0^{\alpha} = \mathbf{m}_0^{\alpha} \otimes \mathbf{n}_0^{\alpha}$ is the Schmid tensor for slip system α, and $\mathbf{m}_0^{\alpha}$ and $\mathbf{n}_0^{\alpha}$ are the slip direction and slip plane normal in the reference configuration, respectively. Finally, the elastic Green–Lagrange strain is defined by

$$\mathbf{E}^e = \frac{1}{2}(\mathbf{C}^e - \mathbf{I}), \tag{3.3}$$

where $\mathbf{C}^e = \mathbf{F}^{e^T}\mathbf{F}^e$ is the elastic right Cauchy–Green deformation tensor.

(ii) Constitutive equations

The second Piola–Kirchhoff stress is related to the elastic Green–Lagrange strain by

$$\mathbf{T}^e = \mathbb{C}[\mathbf{E}^e - (\theta - \theta_0)\mathbf{A}], \tag{3.4}$$

where $\mathbb{C}$ is the fourth-order elastic stiffness tensor, $\mathbf{E}^e$ is the elastic Green–Lagrange strain, $\mathbf{A}$ is the thermal expansion coefficient tensor, θ is the current temperature and θ_0 is a reference temperature. Since tantalum is a cubic material, the thermal expansion tensor is given by $\mathbf{A} = \alpha\mathbf{I}$. Moreover, the elastic stiffness tensor has three unique components and is assumed to degrade with temperature:

$$\left.\begin{aligned} &C_{11}(\theta) = C_{11,0K} - m_{11}\theta, \quad C_{12}(\theta) = C_{12,0K} - m_{12}\theta \\ &C_{44}(\theta) = C_{44,0K} - m_{44}\theta. \end{aligned}\right\} \tag{3.5}$$

The flow rule, incorporating non-Schmid effects, is given by

$$\dot{\gamma}^{\alpha} = \dot{\gamma}_0 \exp\left(-\frac{\Delta G}{k_B\theta}\left\langle 1 - \left\langle \frac{|\tilde{\tau}^{\alpha}| - s^{\alpha}}{\tilde{s}_l} \right\rangle^p \right\rangle^q\right) \quad \text{for } |\tilde{\tau}^{\alpha}| - s^{\alpha} > 0, \tag{3.6}$$

where $\dot{\gamma}_0$ is a reference shear rate, ΔG is the activation energy, k_B is Boltzmann's constant, s^{α} is the slip resistance due to dislocation structure, $\tilde{s}_l$ is the intrinsic lattice resistance, and p and q are exponents that describe the shape of the thermal activation energy barrier. The intrinsic lattice resistance is scaled with temperature as

$$\tilde{s}_l = s_l \frac{\mu(\theta)}{\mu_0}, \tag{3.7}$$

where μ_0 is shear modulus at 0K and $\mu(T)$ is the temperature-dependent anisotropic shear modulus defined by

$$\mu(\theta) = \sqrt{c_{44}(\theta)\frac{c_{11}(\theta) - c_{12}(\theta)}{2}}. \tag{3.8}$$

The resolved shear stress $\tilde{\tau}^\alpha$, accounting for non-Schmid effects, is the stress resolved on to the maximum resolved shear stress plane,

$$\tilde{\tau}^\alpha = (\mathbf{C}^e\mathbf{T}^e) : (\mathbf{S}_0^\alpha + \tilde{\mathbf{S}}_0^\alpha) \approx \mathbf{T}^e : (\mathbf{S}_0^\alpha + \tilde{\mathbf{S}}_0^\alpha), \tag{3.9}$$

where $\tilde{\mathbf{S}}_0^\alpha$ is the non-Schmid tensor, defined as

$$\tilde{\mathbf{S}}_0^\alpha = \sum_{i=1}^{3} \omega_i \tilde{\mathbf{S}}_{0,i}^\alpha. \tag{3.10}$$

The ω_i terms are temperature-dependent weighting factors that determine the strength of the non-Schmid effects. The temperature-dependent weighting factors are given by

$$\omega_i = \omega_{i,ss} + (\omega_{i,0K} - \omega_{i,ss}) \exp(-\theta/\theta_r), \tag{3.11}$$

where θ_r is a reference temperature, and $\omega_{i,0K}$ and $\omega_{i,ss}$ are the weighting factors at 0 K and a saturation value, respectively. In this work, the saturation weighting factor is taken as $\omega_{i,ss} = 0.05 \times \omega_{i,0K}$.

(iii) Hardening Law

The slip resistance is taken to be a modified Taylor Law,

$$s^\alpha = s_0 + \mu b \sqrt{\sum_\beta a^{\alpha\beta} \rho^\beta}, \tag{3.12}$$

where s_0 is a reference slip resistance, μ is the shear modulus, b is the magnitude of the Burgers vector, $a^{\alpha\beta}$ is a slip system interaction matrix and ρ^β is the dislocation density on slip system β. This expression approximates the dislocation interactions that occur between slip systems. The evolution of dislocation density on each slip system is described by the multiplication annihilation law,

$$\dot{\rho}^\alpha = \frac{|\dot{\gamma}_p^\alpha|}{b} \left(\frac{1}{\mathcal{L}^\alpha} - 2y_c^\alpha \sqrt{\rho^\alpha} \right), \tag{3.13}$$

where $\mathcal{L}^\alpha$ is the mean free path of dislocations, given by

$$\frac{1}{\mathcal{L}^\alpha} = \sqrt{\sum_\beta d^{\alpha\beta} \rho^\beta}, \tag{3.14}$$

where $d^{\alpha\beta} = a^{\alpha\beta}/k_1^2$ for self or coplanar interactions and $d^{\alpha\beta} = a^{\alpha\beta}/k_2^2$ for all other interactions. Dislocation annihilation is taken to be in terms of a critical annihilation radius, y_c^α, that is, temperature- and rate-dependent,

$$y_c^\alpha = y_{c0} \left(1 - \frac{k_B\theta}{A_{\text{rec}}} \ln \frac{\dot{\gamma}_p^\alpha}{\dot{\gamma}_0} \right). \tag{3.15}$$

Here, y_{c0} is the reference annihilation capture radius, A_{rec} is the activation energy for recovery, k_B is Boltzmann's constant, θ is the temperature and $\dot{\gamma}_0$ is a reference strain rate.

(b) Proposed factors for stress analysis

Given the model described above, several factors, based on physical knowledge, are proposed as possible candidates that predominantly contribute to the stress conditions at grain boundaries under loading conditions of interest. These factors are derived from both elastic and plastic contributions to represent the heterogeneous deformation in polycrystalline materials [61].

First, the orientation of the grain, described by Euler angles [7,62,63], plays a crucial role in determining the deformation mechanisms and stress distribution within the grain. These Euler angles, notated as θ, Φ and ω, represent a sequence of rotations that transform the crystal coordinate system to the sample coordinate system. The resolved shear stress acting on the slip systems within the grain depends on the grain orientation and the loading direction. This resolved shear stress includes both the classical Schmid factor and the non-Schmid factor [60,64,65], which account for the complex nature of dislocation motion in BCC materials.

In addition to the grain orientation, the elastic properties of the material significantly influence the stress state in polycrystalline materials through two main mechanisms. First, the elastic constants, represented by the fourth-order tensor $\mathbb{C}$, vary with crystallographic orientation due to material anisotropy. Second, the elastic strain $\mathbf{E}^e$, which represents reversible deformation, is related to stress through these orientation-dependent constants. Together, these factors lead to incompatible elastic deformation at grain boundaries [15], creating stress concentrations and discontinuities at interfaces even under uniform loading conditions. Therefore, both elastic constants in the global coordinate system and elastic strain variations across grain boundaries must be considered when analysing stress distributions in polycrystalline materials.

Next, analogous to the elastic constants, these stress disparities can cause mismatched plastic flow across the boundaries. This effect is particularly pronounced in BCC materials where plastic deformation is dominated by screw dislocations. These screw dislocations exhibit unique characteristics, dissociating into three partial dislocations at rest and forming an equidistant triad within the crystallographic lattice [66,67]. Their non-planar core structure prevents accurate representation through classical Schmid rules [68,69] and significantly reduces dislocation mobility compared to edge dislocations. Therefore, when considering the current state of plastic deformation in each grain, we include both classical Schmid factors and non-Schmid effects (equation (3.9)) in our normalized resolved shear stresses. These 48 non-Schmid factors are sorted by magnitude and re-termed $\hat{\tau}_I$, where $I = 1$ indicates the largest factor. The first five largest factors are included in our analysis to capture the primary deformation mechanisms and their contribution to stress heterogeneity at grain boundaries.

Moreover, the total statistically stored dislocation density

$$\rho_{\text{ssd}} \equiv \sum_{\alpha} \rho^{\alpha} \tag{3.16}$$

is included as this is a quantity that is readily available to both single-crystal models and isotropic models.

To incorporate the current state of plastic deformation, the spatial strain measure is considered:

$$\mathbf{B}^p\mathbf{B}^p \equiv \mathbf{F}^p\mathbf{F}^{p^T} = \sum_{i=1}^{3} \lambda_i^{p^2} \mathbf{l}_i^p \otimes \mathbf{l}_i^p. \tag{3.17}$$

Its principal stretches and directions are computed. The rate of plastic deformation, provided by the symmetric part of the plastic velocity gradient,

$$\mathbf{D}^p = \text{sym}\, \mathbf{L}^p = \sum_{i=1}^{3} \lambda_i^p \mathbf{v}_i^p \otimes \mathbf{v}_i^p, \tag{3.18}$$

is also taken into account. The misorientation between the accumulated plastic flow directions and the plastic slip rate directions between grains is considered. The rate of plastic deformation is captured by $\mathbf{D}^p$ in equation (3.18), with eigenvalues λ_i^p indicating deformation rates

Table 1. Microstructural factors considered in the candidate function library. Notations with superscript max indicate quantities evaluated using the hotspot local information.

notation	description
θ, Φ, ω	grain orientation relative to sample coordinates.
$\sqrt{\rho_{\text{ssd}}}$	square root of the mean of statistically stored dislocation density for all grains as shown in equation (3.16).
$\hat{\tau}_{i,Gn}(\hat{\tau}^{\max}_{i,Gn})$	the *i*th largest compressive non-Schmid factors in grain *n*.
$\mathcal{C}_{ij,Gn}$	the *ij*th component of the elastic constants for grain *n*, where *i*, *j* represent the row and column coordinates in a 6 × 6 matrix. The matrix, using Voigt notation, compactly describes the full elastic tensor.
$\mathbf{E}^{e}_{ij,Gn}$	the *ij*th component of the grain-volume-averaged strain in grain *n* similar to $\mathcal{C}_{ij,Gn}$.
$\mathbf{v}_{i,Gn}\cdot\mathbf{v}_{j,Gm}(\mathbf{v}^{\max}_{i,Gn}\cdot\mathbf{v}^{\max}_{j,Gm})$	the misorientation of the principal stretch rate directions (i.e. eigenvectors) of $\mathbf{D}^p$ in equation (3.18) corresponding to the *i*th eigenvalue of grain *n* with the *j*th eigenvalue of grain *m*.
$\lambda_{i,Gn}(\lambda^{\max}_{i,Gn})$	the *i*th eigenvalue (sorted in descending order) of $\mathbf{D}^p$ for grain *n*.
$\mathbf{u}_{i,Gn}\cdot\mathbf{u}_{j,Gm}(\mathbf{u}^{\max}_{i,Gn}\cdot\mathbf{u}^{\max}_{j,Gm})$	similar to $\mathbf{v}_{i,Gn}\cdot\mathbf{v}_{j,Gm}$, but computed from $\mathbf{B}^p$ in equation (3.17).
$\mu_{i,Gn}(\mu^{\max}_{i,Gn})$	similar to $\lambda_{i,Gn}$, but computed from $\mathbf{B}^p$.

and eigenvectors $\mathbf{v}^p_i$ describing deformation directions. These measures are particularly important when considering grain-boundary interactions, as demonstrated by experimental observations [12].

To summarize the above justifications, the candidate factors are listed in table 1.

4. Experimental set-up

A series of idealized microstructure configurations with increasing complexity are simulated to identify specific physical features leading to high-stress spots in deformed polycrystalline metals. All simulations are conducted at a strain rate of $\dot{\epsilon} = 10^5\ \text{s}^{-1}$, a rate commonly observed in dynamic loading conditions known to cause damage in tantalum.

The study begins with a single-crystal configuration used to validate the statistical analysis methodology, for which 200 simulations are performed to establish adequate statistical sampling. This is followed by two distinct bi-crystal configurations, each requiring 800 simulations: a cylinder with a grain boundary perpendicular to the loading direction and another with a grain boundary parallel to the loading direction. In addition, a quad-crystal configuration is simulated to incorporate both perpendicular and parallel grain-boundary interactions with 2000 simulations. Finally, an octu-crystal (e.g. eight-grain) configuration is implemented to approach realistic microstructure complexity, with 3000 simulations to fully characterize the system behaviour. Note that the increased number of simulations is utilized to ensure the increased complexity of the model still provides statistically significant results under the KS test.

5. Analysis of grain-boundary effects on stress distributions

(a) Single-crystal analysis: establishing the baseline stress state

To establish a fundamental understanding of stress states, an analysis of single-crystal data is first conducted. For the single-crystal case, the resultant stress (rather than maximum stress) is used, defined as the volume-averaged stress over the entire crystal domain, since there are no

Table 2. Causation entropy analysis on the single-crystal configuration.

feature	causation entropy	feature	causation entropy
$\mathbf{E}^e_{33}$	0.3108	$\mathbf{E}^e_{11}$	0.0013
$\mathcal{C}_{33}$	0.2135	$\mathcal{C}_{11}$	0.0003
$\sqrt{\rho_{\mathrm{ssd}}}$	0.0753	$\mathbf{E}^e_{12}$	0.0002
$\hat{\tau}_4$	0.0091	$\hat{\tau}_2$	0.0002
$\mathbf{E}^e_{23}$	0.0070	ϕ	0.0001
$\hat{\tau}_3$	0.0058	$\mathcal{C}_{22}$	7.98×10^{-5}
$\hat{\tau}_5$	0.0056	$\mathbf{E}^e_{13}$	1.97×10^{-5}
θ	0.0025	ω	1.57×10^{-5}
$\mathbf{E}^e_{22}$	0.0023	$\hat{\tau}_1$	1.02×10^{-5}

internal interfaces that could lead to stress concentrations. This approach allows for the isolation of intrinsic material behaviour before introducing the complexities of grain boundaries and multi-grain interactions. Two hundred single-crystal simulations are performed, varying initial orientations to capture various possible microstructural states. The resulting stress data then facilitate effective statistical analysis to identify key relationships and trends.

The analysis is initiated with a linear combination of a subset of the features to establish a baseline and identify the most influential variables. Causation entropy calculations are used to select features that are most relevant to stress prediction. The causation entropies for all features are listed in table 2.

The causation entropy analysis revealed several key insights into the relative importance of different microstructural features in predicting stress states. As evident from table 2, $\mathbf{E}^e_{33}$, $\mathcal{C}_{33}$ and $\sqrt{\rho_{\mathrm{ssd}}}$ exhibit the highest causation entropy values, which implies their significant influence on the stress state. This is consistent with physical intuition since these parameters represent critical aspects of the material's elastic behaviour, crystal structure and dislocation density. Interestingly, while the non-Schmid factors $\hat{\tau}_i$ show varying degrees of causation entropy, they collectively contribute substantial information to stress prediction. Although $\hat{\tau}_1$ shows a relatively low individual causation entropy, all five non-Schmid factors are included in the model to maintain a comprehensive representation of the non-Schmid effects.

Based on these insights, a simple linear relationship between the stress state and the most influential variables is obtained:

$$\sigma_{\mathrm{model}} = \beta_0\sqrt{\rho_{\mathrm{ssd}}} + \beta_1\mathbf{E}^e_{33} + \beta_2\mathcal{C}_{33} + \sum_{i=1}^{5}\beta_{3i}\hat{\tau}_i, \tag{5.1}$$

where σ_{model} is the predicted resultant stress, ρ_{ssd} is the statistically stored dislocation density, $\mathcal{C}_{33}$ is an elastic constant, $\mathbf{E}^e_{33}$ is an average strain, $\hat{\tau}_i$ are the non-Schmid factors and β_0, β_1, β_2 and β_{3i} are the model coefficients.

The residual of the identified system in equation (5.1) is then calculated by taking the difference between σ_{true} and σ_{model}, the statistics of which reflect the uncertainty in such a relationship. Figure 3 shows the PDFs of the resultant stress and the associated residual. The results reveal several important aspects of the model's performance. Notably, the variance of the residual is only 1.27% of the original resultant stress. Furthermore, the skewness and kurtosis of the residual distribution are close to 0 and 3, respectively, indicating that the residual closely approximates a Gaussian distribution. This near-Gaussian nature of the small residual suggests that it behaves similarly to white noise, implying that the linear model of equation (5.1) has successfully captured the significant features of the resultant stress distribution. The stark contrast between the two

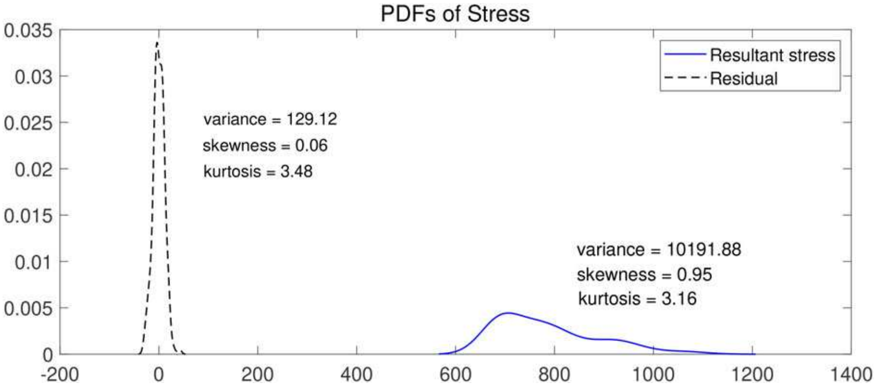

Figure 3. PDFs of the resultant stress (blue line) and residual (dashed black line) from the linear regression model in equation (5.1) for single-crystal configurations.

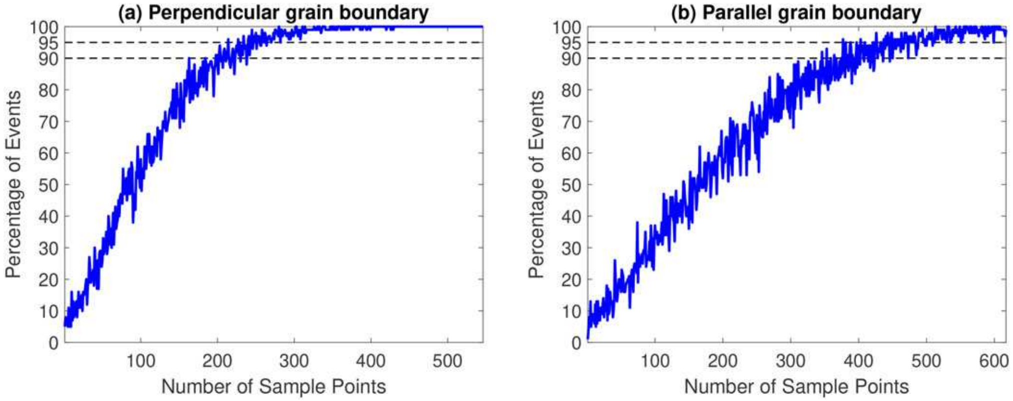

Figure 4. KS tests results on the maximum stress if it occurred near a grain boundary in a bi-crystal configuration. Panels (a) and (b) show the results for the simulations with perpendicular grain boundary and parallel grain boundary, respectively. For each number of sample points, there are 100 times KS tests to record the percentage of passing tests.

distributions demonstrates the effectiveness of the linear relationship in explaining the majority of the stress variation in single-crystal configurations.

(b) Bi-crystal analysis: introducing grain-boundary effects

This subsection investigates two primary bi-crystal configurations: one with the grain-boundary plane perpendicular to the loading direction and another with the grain-boundary plane parallel to the loading direction. These idealized set-ups provide fundamental insights into grain-boundary effects while maintaining a tractable level of complexity.

For each configuration, 2000 simulations with different crystallographic orientations are conducted. Among these, approximately 600 simulations exhibit maximum von Mises stress near the grain boundary, which is used for the feature analysis carried out below. To ensure the statistical robustness of the results, KS tests are applied to the subset where maximum stress occurs at grain boundaries. Figure 4 illustrates these results, with both types reaching the threshold rejection rate of 95%. This high rejection rate confirms not only the statistical significance of the sampling but also the non-Gaussian nature of the stress distributions, justifying the need for advanced statistical techniques in subsequent analyses.

Initial investigations employed linear models to identify critical factors influencing stress states. However, these linear models are insufficient in capturing the full complexity of the stress distributions (not shown here), particularly the non-Gaussian features identified by the KS tests. Given the limitations of linear models and guided by physical knowledge of crystal plasticity, a general nonlinear analysis approach is adopted. The candidate feature function library for the nonlinear analysis includes:

- **Linear terms**: These represent the direct, first-order effects of various factors as listed in table 1 on stress, providing a baseline for the model.
- **Quadratic combinations of stretch rates $\boldsymbol{\lambda}_{i,Gn}$ and $\boldsymbol{\lambda}^{\mathbf{max}}_{i,Gn}$**: These terms represent the nonlinear effects of the rate of plastic deformation on stress states, particularly important in dynamic loading conditions.
- **Squares of $\mathbf{u}_{i,Gn} \cdot \mathbf{u}_{j,Gm}$ and $\mathbf{u}^{\mathbf{max}}_{i,Gn} \cdot \mathbf{u}^{\mathbf{max}}_{j,Gm}$**: These terms represent the nonlinear effects of grain misorientation on stress states, particularly at grain boundaries.
- **Quadratic combinations of cumulative stretch $\boldsymbol{\mu}_{i,Gn}$ and $\boldsymbol{\mu}^{\mathbf{max}}_{i,Gn}$**: These terms capture the nonlinear effects of accumulated plastic deformation on stress states, particularly significant in cases of large plastic strains.
- **Squares of $\mathbf{v}_{i,Gn} \cdot \mathbf{v}_{j,Gm}$ and $\mathbf{v}^{\mathbf{max}}_{i,Gn} \cdot \mathbf{v}^{\mathbf{max}}_{j,Gm}$**: These terms represent the nonlinear effects of grain misorientation on stress states, particularly at grain boundaries.
- **Quadratic combinations of elastic constants $\boldsymbol{\mathcal{C}}_{ii,Gn}$**: These terms model the complex local elastic behaviour, especially stress concentrations arising from mismatches in elastic properties between adjacent grains.
- **Quadratic combinations of average strain $\mathbf{E}^{e}_{ii,Gn}$**: Analogous to the elastic constant terms, these quadratic combinations are used to capture nonlinear effects in the average strain state of each grain, providing a more comprehensive representation of the strain–stress relationship at the grain level.
- **Quadratic combinations of non-Schmid factors $\boldsymbol{\tau}_{i,Gn}$ and $\boldsymbol{\tau}^{\mathbf{max}}_{i,Gn}$**: These terms represent the nonlinear effects of non-Schmid behaviour.

Introducing quadratic terms necessitates further categorizing of these features, as their physical significance extends beyond individual contributions. Instead of treating each quadratic term individually, several categories are divided to reflect their underlying physical interactions. This categorization is applied not only for simplification but also for showing deeper insights into the mechanisms driving stress distribution in bi-crystal cylinder simulations. Here, $\mathbf{E}^{e}_{ij,Gn}$ is taken as an illustrative example; their quadratic combinations can be divided into two groups: (i) intragrain interactions and (ii) intergrain interactions. The former represents the quadratic combinations of elastic average strain components within the same grain, while the latter means the quadratic combinations of those components between different grains, which describe the complex interplay of elastic strains across the grain boundary. From a physical standpoint, the intergrain interactions are of particular interest as they directly relate to the stress state at and near the grain boundary, a critical region for potential damage initiation. This emphasis on intergrain interactions is not merely theoretical but is substantiated by our causation entropy analysis.

Figure 5 presents a heatmap of the causation entropy for various combinations of elastic average strain tensor components in the bi-crystal cylinder under the perpendicular grain boundary. The colour and size of each circle represent the magnitude of causation entropy, with larger, warmer-coloured circles indicating higher values. Notably, the intergrain interactions (upper-right quadrant) consistently exhibit higher causation entropy values compared to the intragrain interactions (upper-left quadrant and lower-right quadrant). This visual representation clearly shows that this intergrain interaction has a more significant information contribution compared to intragrain interactions. Such a finding aligns with physical understanding: the interface between grains, where material properties can change abruptly, is likely to be a key determinant of the system's stress distribution.

In summary, features considering all quadratic combinations can be generally categorized into intragrain and intergrain interactions. For some features, there is no physical meaning in different

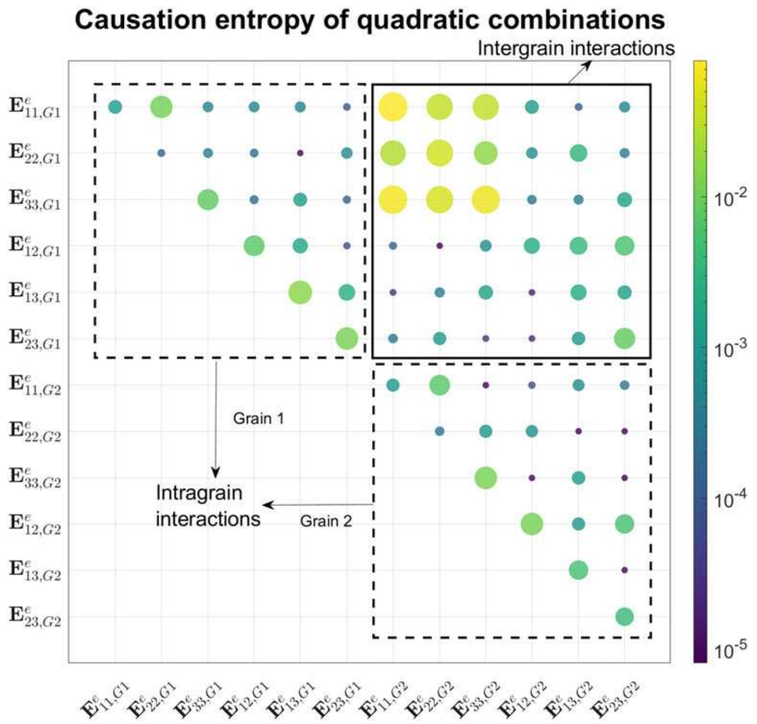

Figure 5. Heatmap of causation entropy for quadratic combinations of elastic strain tensor components ($\mathbf{E}_{ij,Gn}$) in a bi-crystal system. The colour and size of each circle represent the magnitude of causation entropy. The dashed boxes highlight intragrain interactions (the upper-left is for grain 1 and the lower-right is for grain 2) and intergrain interactions (upper-right).

sub-variables with different grains. Therefore, the intragrain interaction only contains its square terms. To formalize this categorization, a notation system is introduced for generic features $\{\mathcal{X}\}$:

— $\{\mathcal{X}\}_L$: linear terms;
— $\{\mathcal{X}\}_S$: quadratic terms representing intragrain interactions or square terms;
— $\{\mathcal{X}\}_C$: quadratic terms representing intergrain interactions.

Figure 6 illustrates the causation entropy values for various categories under two distinct grain-boundary configurations: perpendicular and parallel. The categories are arranged in descending order based on the mean causation entropy value across both configurations.

The selection of candidate features extends beyond purely statistical considerations to ensure the comprehensive representation of all relevant physical mechanisms. The foundation of the model begins with the dislocation density term $\sqrt{\rho_{\rm ssd}}$. Although this term shows relatively low causation entropy in the bi-crystal configuration, it is retained as it represents the fundamental carrier of plastic deformation and provides continuity with single-crystal behaviour. In addition, it is noteworthy that the number of terms within each category may significantly influence the magnitude of the causation entropy during the dislocation density $\sqrt{\rho_{\rm ssd}}$ category. Actually, the dislocation density $\sqrt{\rho_{\rm ssd}}$ shows the high causation entropy when computing them for individual features. The linear elastic terms $\mathbf{E}^e_{ij,Gn}$ and $\mathcal{C}_{ii,Gn}$ are included as they describe the basic elastic response, which is essential for maintaining mechanical equilibrium regardless of their causation entropy values.

The linear elastic terms $\mathbf{E}^e_{ij,Gn}$ and $\mathcal{C}_{ii,Gn}$ are included as fundamental descriptors of the primary elastic response, essential for maintaining mechanical equilibrium regardless of their causation entropy values. The cross products of elastic constants $\mathcal{C}_{ii,Gn}\mathcal{C}_{ii,Gm}$ and elastic strains $\mathbf{E}^e_{ij,Gn}\mathbf{E}^e_{kl,Gm}$ are incorporated to characterize elastic incompatibility at grain boundaries. In addition, square

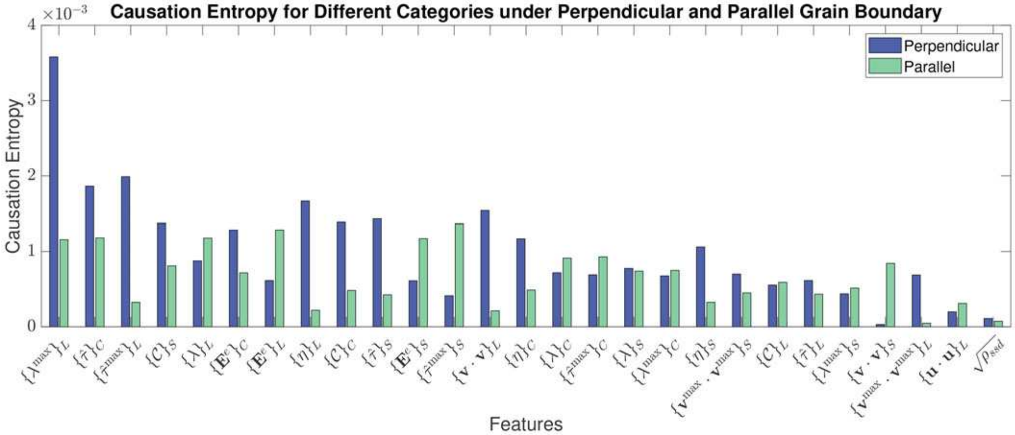

Figure 6. Causation entropies computed for different categories for maximum stress under perpendicular grain boundary (dark blue) and parallel grain boundary (green). The *x*-axis label represents the different categories, where the subscripts indicate the linear terms, square terms and intergrain interaction terms.

terms $(\mathcal{C}_{ij,Gn})^2$ and $(\mathbf{E}^e_{ij,Gn})^2$ are included to capture intergrain interactions and local nonlinear effects. Notably, all these quadratic terms demonstrate high causation entropy values in the analysis. These terms naturally emerge in expressions involving differences between grain features, taking the form $(X_{In} - X_{Jm})^2$, where X represents various physical quantities.

Non-Schmid factor interactions play a crucial role in our model, as evidenced by their high causation entropy values. The terms $\hat\tau_{i,Gn}\hat\tau_{j,Gm}$ and those evaluated using the local hotspot information are particularly important as they capture the complex interactions between different slip systems that can lead to stress concentrations at grain boundaries. In addition, the orientation-dependent terms $(\mathbf{v}_{i,Gn}\cdot\mathbf{v}_{i,Gm})^2$ and the corresponding terms evaluated using the local hotspot information describe misorientation effects, which, while showing moderate causation entropy, are essential for understanding stress development at grain boundaries.

Based on the above considerations, the resulting statistical model is written as

$$\begin{aligned}\sigma_{\text{model}} = \beta_0\sqrt{\rho_{\text{ssd}}} &+ \sum_{n}^{N_{\text{gr}}}\sum_{i=1}^{3}\beta_{1ni}\lambda_{i,Gn} + \sum_{n}^{N_{\text{gr}}}\sum_{i=1}^{3}\beta_{2ni}\lambda^{\max}_{i,Gn} + \sum_{n}^{N_{\text{gr}}}\sum_{i,j=1}^{3}\beta_{3nij}\mathbf{E}^e_{ij,Gn}\\ &+ \sum_{n}^{N_{\text{gr}}}\sum_{i=1}^{3}\beta_{4ni}\mathcal{C}_{ii,Gn} + \sum_{m>n}^{N_{\text{gr}}}\sum_{i,j=1}^{3}\beta_{5nmij}\mathbf{E}^e_{ij,Gn}\mathbf{E}^e_{ij,Gm} + \sum_{m>n}^{N_{\text{gr}}}\sum_{i=1}^{3}\beta_{6nmi}\mathcal{C}_{ii,Gn}\mathcal{C}_{ii,Gm}\\ &+ \sum_{n}^{N_{\text{gr}}}\sum_{i,j=1}^{3}\beta_{7nij}(\mathbf{E}^e_{ij,Gn})^2 + \sum_{n}^{N_{\text{gr}}}\sum_{i=1}^{3}\beta_{8nij}(\mathcal{C}^e_{ii,Gn})^2 + \sum_{m>n}^{N_{\text{gr}}}\sum_{i,j=1}^{5}\beta_{9nmij}\hat\tau_{i,Gn}\hat\tau_{j,Gm}\\ &+ \sum_{m>n}^{N_{\text{gr}}}\sum_{i=1}^{5}\beta_{10nmi}\hat\tau^{\max}_{i,Gn}\hat\tau^{\max}_{i,Gm} + \sum_{n}^{N_{\text{gr}}}\sum_{i=1}^{3}\beta_{11ni}(\mathbf{v}_{i,Gn}\cdot\mathbf{v}_{i,Gm})^2\sum_{n}^{N_{\text{gr}}}\sum_{i=1}^{3}\beta_{12ni}(\mathbf{v}^{\max}_{i,Gn}\cdot\mathbf{v}^{\max}_{i,Gm})^2,\end{aligned} \tag{5.2}$$

where σ_{model} is the predicted maximum stress and β represents the coefficients for the different terms.

To investigate the model accuracy, the residual σ_{res} is computed between the true maximum stress and the predicted values from the above statistical model, $\sigma_{\text{res}} = \sigma_{\text{truth}} - \sigma_{\text{model}}$. Figure 7 shows such results. For both bi-crystal configurations, the residual (black lines) has a significantly smaller variance, less than 10%, than the truth (red-lines), which implies the validity of the feature analysis in equation (5.2).

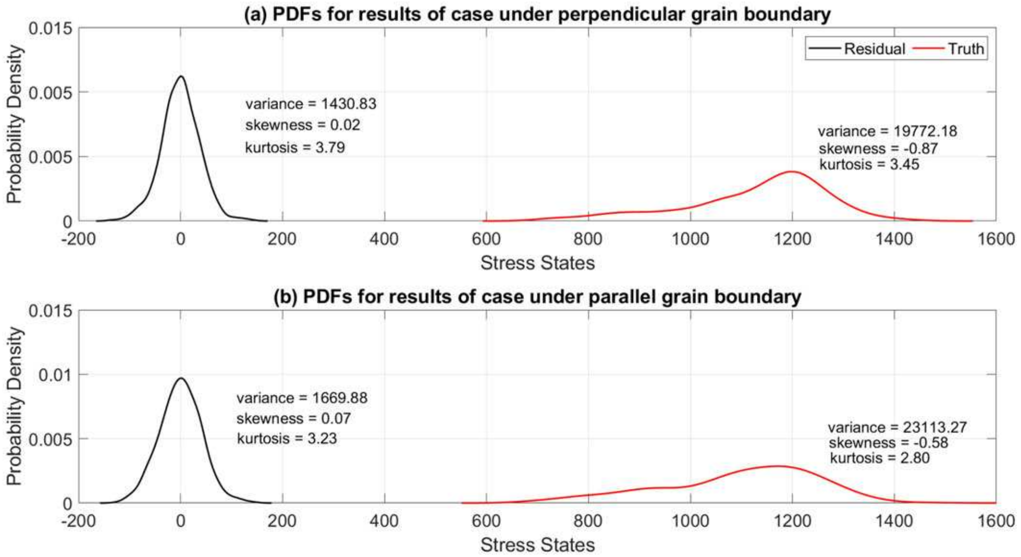

Figure 7. PDFs of the true maximum stress (red lines) and the residual (black lines). (a) The results for the bi-crystal cylinder under perpendicular grain boundary. (b) The results for the bi-crystal configuration under parallel grain boundary.

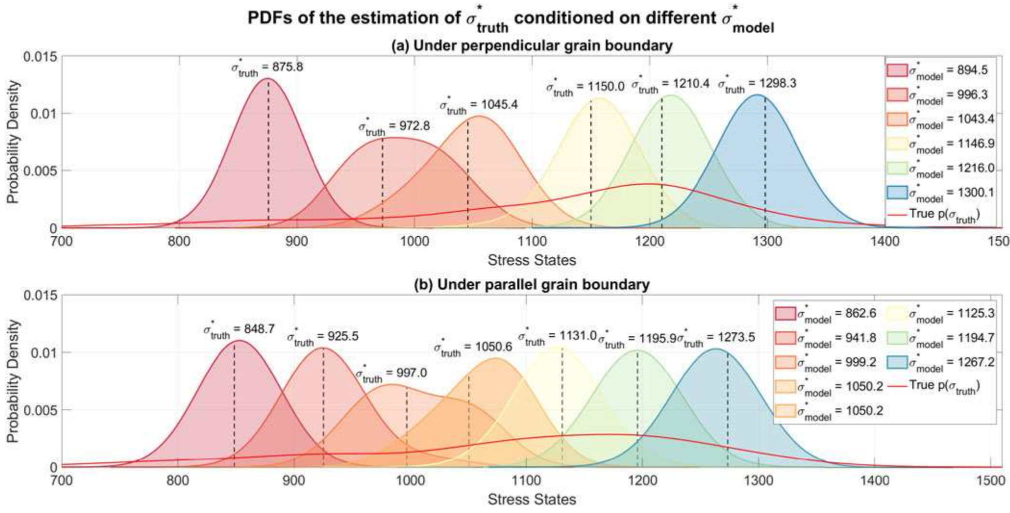

Figure 8. PDFs of the predicted maximum stress σ^*_{truth} conditioned on different values of σ^*_{model} and the true maximum stress distribution $p(\sigma_{\text{truth}})$. (a) The PDFs of the estimation of σ^*_{truth} under perpendicular grain boundary. (b) The PDFs of the estimation of σ^*_{truth} under parallel grain boundary. Each colour corresponds to a different conditioning value of σ^*_{model}. The vertical dashed lines indicate the position of true σ^*_{truth} for each predicted PDF.

Figure 8 presents the forecast of the stress value with the associated uncertainty. Here, the forecast is based on the CGMM. It is seen that the forecast uncertainty varies as the predicted value of the maximum stress σ^*_{model}. The uncertainty is not necessarily a Gaussian distribution, which implies the necessity of using the mixture presentations. Due to the additional information of the conditioned state, i.e. the predicted σ^*_{model}, the uncertainty in the prediction is usually slightly smaller than the total residual shown in figure 7, which provides more confident forecast results.

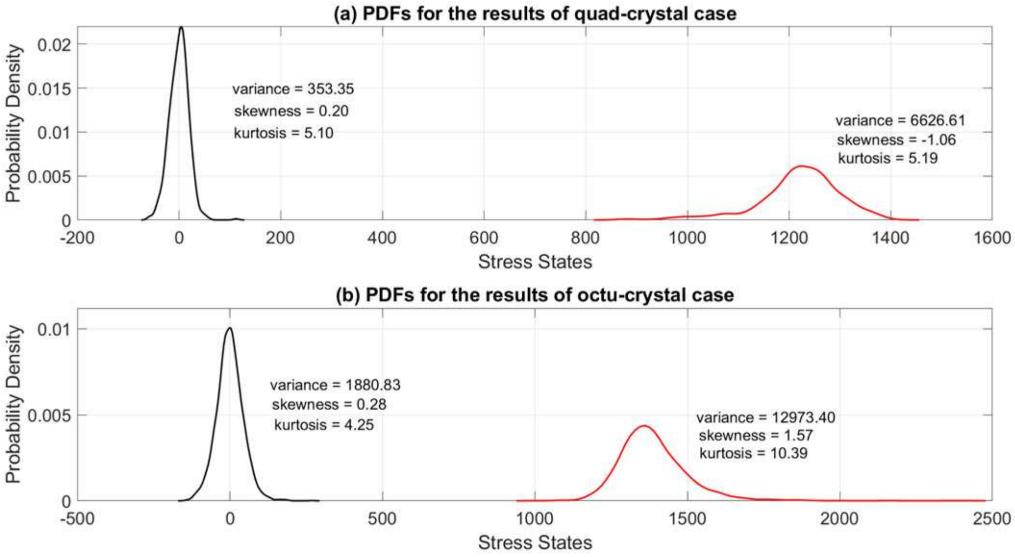

Figure 9. PDFs of true maximum stress distribution (red lines) and model residuals (black lines) in multi-grain configurations. (a) The results for quad-crystal cylinder configuration. (b) The results for octu-crystal cylinder configuration.

(c) Multi-grain configurations

Following the statistical model developed from bi-crystal analysis, more complex configurations containing multiple grain boundaries are investigated to examine the model's applicability to realistic polycrystalline structures. These multi-grain configurations, including quad-crystal and octu-crystal arrangements, serve as intermediate steps between idealized bi-crystal cases and actual polycrystalline materials.

The statistical model is adapted for these multi-grain configurations to maintain computational efficiency while preserving essential physical mechanisms. As the number of grains increases, the full model from the bi-crystal case would include an overwhelming number of interaction terms, scaling approximately as $O(N_g^2)$ for N_g grains. Therefore, a simplified version, equation (5.3), is employed, retaining only the terms with high causation entropy values identified through previous analysis. This refined model preserves fundamental terms, including dislocation density, elastic constants, elastic strains and critical local features while selectively including the most significant intergrain interactions. The resulting statistical reduced-order model reads

$$\begin{aligned}\sigma_{\text{model}} = {} & \beta_0\sqrt{\rho_{\text{ssd}}} + \sum_{n}^{N_{\text{gr}}}\sum_{i=1}^{3}\beta_{1ni}\lambda_{i,Gn} + \sum_{n}^{N_{\text{gr}}}\sum_{i=1}^{3}\beta_{2ni}\lambda_{i,Gn}^{\max} + \sum_{n}^{N_{\text{gr}}}\sum_{i,j=1}^{3}\beta_{3nij}\mathbf{E}^e_{ij,Gn} + \sum_{n}^{N_{\text{gr}}}\sum_{i=1}^{3}\beta_{4ni}\mathcal{C}_{ii,Gn} \\ & + \sum_{m>n}^{N_{\text{gr}}}\sum_{i,j=1}^{3}\beta_{5nmij}\mathbf{E}^e_{ij,Gn}\mathbf{E}^e_{ij,Gm} + \sum_{m>n}^{N_{\text{gr}}}\sum_{i=1}^{3}\beta_{6nmi}\mathcal{C}_{ii,Gn}\mathcal{C}_{ii,Gm} + \sum_{n}^{N_{\text{gr}}}\sum_{i,j=1}^{3}\beta_{7nij}(\mathbf{E}^e_{ij,Gn})^2 \\ & + \sum_{n}^{N_{\text{gr}}}\sum_{i=1}^{3}\beta_{8nij}(\mathcal{C}^e_{ii,Gn})^2 + \sum_{m>n}^{N_{\text{gr}}}\sum_{i=1}^{3}\beta_{9nmi}\hat{\tau}^{\max}_{i,Gn}\hat{\tau}^{\max}_{i,Gm} + \sum_{n}^{N_{\text{gr}}}\sum_{i=1}^{3}\beta_{10ni}(\mathbf{v}^{\max}_{i,Gn}\cdot\mathbf{v}^{\max}_{i,Gm})^2. \end{aligned} \tag{5.3}$$

Similar to the cases for the configurations of bi-crystal cylinders, figure 9 compares the PDFs of the true maximum stress (red lines) and the residual of the statistically resulting model (black lines). The residual distribution has a much smaller variance than the truth. It also remains approximately Gaussian, suggesting the model successfully captures the primary sources of non-Gaussian behaviour in the stress field. Similarly, figure 10 shows the forecast and

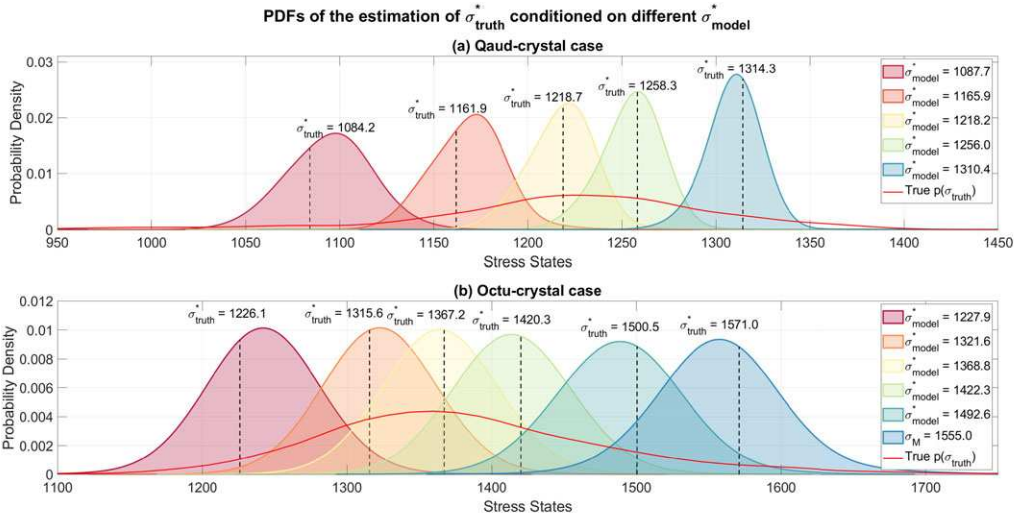

Figure 10. Similar to figure 8, but for the quad-crystal case and octu-crystal case.

its uncertainties. The forecast mean values are close to the truth, while the uncertainty is again smaller than the total residual overall, indicating the necessity of using the CGMM. The successful application of this streamlined model to more complex arrangements suggests that the fundamental physics of grain-boundary stress concentrations can be captured with a reduced set of carefully selected interaction terms. Notably, despite their increasing geometric complexity, the model's performance remains consistent across both quad-crystal and octu-crystal configurations. This robustness in performance across different levels of microstructural complexity validates the selection of essential interaction terms in the simplified model, which suggests the potential applicability of the statistically reduced-order modelling framework to even larger polycrystalline systems.

6. Discussion and conclusion

The framework developed in this study advances our ability to analyse complex material behaviour by uniquely combining physical principles with statistical learning. Traditional approaches to studying the damage behaviour of polycrystalline metallic materials rely on computationally expensive full-field simulations or homogenized engineering models, which sacrifice physical accuracy for computational efficiency. Most existing analysis tools also have difficulties in representing the non-Gaussian feature of extreme events, which is, however, crucial for damage prediction. This framework avoids the above limitations through three key components.

The uniqueness of the proposed framework consists first of a systematic way of distinguishing between causal relationships and mere correlations. Unlike traditional correlation-based methods, the causation entropy analysis identifies true underlying causal relationships between microstructural features and stress states. This distinction is crucial because it identifies physically meaningful mechanisms and enables principled model reduction by retaining only features with genuine causal influence. This provides a mathematical base for experimentally observed damage patterns in future work.

Beyond stress analysis in polycrystals, this framework establishes a general methodology for studying complex material phenomena. Its key strength is demonstrated by successful extrapolation from simple to complex systems, and models trained on bi-crystal configurations accurately predict stress states in quad-crystal and octu-crystal systems. This robustness suggests that the identified physical mechanisms are fundamental rather than specific to particular configurations, which makes the framework valuable for analysing realistic material systems.

The framework is also efficient because it combines physical insight and statistical techniques. Traditional crystal plasticity simulations become prohibitively expensive for complex polycrystalline systems. Using causation entropy to identify essential features and then employing CGMMs for probabilistic predictions reduces computational time by orders of magnitude while maintaining accuracy in predicting extreme stress states. In fact, once the statistical model is developed, the forecast, which relies only on computing the few selected features without running the original computational models, becomes highly efficient. In addition, the incorporation of CGMM for uncertainty quantification provides probabilistic predictions at a low computational cost. This efficiency enables the rapid evaluation of multiple material configurations, which is critical for practical materials design applications.

These advances establish a new framework for analysing material behaviour, achieving accuracy and efficiency. As materials science increasingly confronts complex, hierarchical systems, such physics-informed statistical approaches will become essential for bridging scales and understanding structure-property relationships. Future developments could extend this methodology to dynamic loading conditions and integration with experimental data, further broadening its effect across materials science applications.

Data accessibility. The parameters of the statistical models presented in equations (5.1)–(5.3) and the CGMM for different cases are provided in the electronic supplementary material [70].

Declaration of AI use. We have not used AI-assisted technologies in creating this article.

Authors' contributions. Y.Z.: formal analysis, investigation, methodology, software, visualization, writing—original draft; S.D.D.: data curation, formal analysis, investigation, methodology, software; C.A.B.: conceptualization, funding acquisition, project administration, supervision, writing—review and editing; N.C.: conceptualization, funding acquisition, project administration, supervision, writing—original draft.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. C.A.B. and N.C. are grateful for the support from NSF DMREF-CMMI 2118399. S.D.D. and Y.Z. are supported as research assistants under this grant.

References

1. Czarnota C, Mercier S, Molinari A. 2006 Modelling of nucleation and void growth in dynamic pressure loading, application to spall test on tantalum. *Int. J. Fract.* **141**, 177–194. (doi:10.1007/s10704-006-0070-y)
2. Gray GT, Bourne N, Livescu V, Trujillo CP, MacDonald S, Withers P. 2014 The influence of shock-loading path on the spallation response of Ta. *J. Phys: Conf. Ser.* **500**, 112031. (doi:10.1088/1742-6596/500/11/112031)
3. Wilkerson J, Ramesh K. 2016 A closed-form criterion for dislocation emission in nano-porous materials under arbitrary thermomechanical loading. *J. Mech. Phys. Solids* **86**, 94–116. (doi:10.1016/j.jmps.2015.10.005)
4. Wilkerson J. 2017 On the micromechanics of void dynamics at extreme rates. *Int. J. Plast.* **95**, 21–42. (doi:10.1016/j.ijplas.2017.03.008)
5. Versino D, Bronkhorst CA. 2018 A computationally efficient ductile damage model accounting for nucleation and micro-inertia at high triaxialities. *Comput. Methods Appl. Mech. Eng.* **333**, 395–420. (doi:10.1016/j.cma.2018.01.028)
6. Nguyen T, Luscher DJ, Wilkerson JW. 2019 The role of elastic and plastic anisotropy in intergranular spall failure. *Acta Mater.* **168**, 1–12. (doi:10.1016/j.actamat.2019.01.033)
7. Bronkhorst C, Cho H, Marcy P, Vander Wiel S, Gupta S, Versino D, Anghel V, Gray III G. 2021 Local micro-mechanical stress conditions leading to pore nucleation during dynamic loading. *Int. J. Plast.* **137**, 102903. (doi:10.1016/j.ijplas.2020.102903)
8. Fensin SJ, Valone SM, Cerreta EK, Gray GT. 2012 Influence of grain boundary properties on spall strength: grain boundary energy and excess volume. *J. Appl. Phys.* **112**, 083529. (doi:10.1063/1.4761816)

9. Fensin S, Valone S, Cerreta E, Escobedo-Diaz J, Gray G, Kang K, Wang J. 2012 Effect of grain boundary structure on plastic deformation during shock compression using molecular dynamics. *Modell. Simul. Mater. Sci. Eng.* **21**, 015011. (doi:10.1088/0965-0393/21/1/015011)
10. Aifantis KE. 2009 Interfaces in crystalline materials. *Procedia Eng.* **1**, 167–170. (doi:10.1016/j.proeng.2009.06.039)
11. Ashmawi W, Zikry M. 2003 Grain boundary effects and void porosity evolution. *Mech. Mater.* **35**, 537–552. (doi:10.1016/S0167-6636(02)00269-7)
12. Ziegler A, Campbell G, Kumar M, Stölken J. 2003 Effects of grain boundary constraint on the Constitutive response of tantalum bicrystals. *MRS Online Proc. Lib.* **779**, 641–646. (doi:10.1557/PROC-779-W6.4)
13. Hahn EN, Fensin SJ, Germann TC, Gray III GT. 2018 Orientation dependent spall strength of tantalum single crystals. *Acta Mater.* **159**, 241–248. (doi:10.1016/j.actamat.2018.07.073)
14. Weaver JS, Jones DR, Li N, Mara N, Fensin S, Gray III GT. 2018 Quantifying heterogeneous deformation in grain boundary regions on shock loaded tantalum using spherical and sharp tip nanoindentation. *Mater. Sci. Eng.: A* **737**, 373–382. (doi:10.1016/j.msea.2018.09.075)
15. Lebensohn RA, Escobedo JP, Cerreta EK, Dennis-Koller D, Bronkhorst CA, Bingert JF. 2013 Modeling void growth in polycrystalline materials. *Acta Mater.* **61**, 6918–6932. (doi:10.1016/j.actamat.2013.08.004)
16. Lieberman EJ, Lebensohn RA, Menasche DB, Bronkhorst CA, Rollett AD. 2016 Microstructural effects on damage evolution in shocked copper polycrystals. *Acta Mater.* **116**, 270–280. (doi:10.1016/j.actamat.2016.06.054)
17. Francis T *et al.* 2021 Multimodal 3D characterization of voids in shock-loaded tantalum: implications for ductile spallation mechanisms. *Acta Mater.* **215**, 117057. (doi:10.1016/j.actamat.2021.117057)
18. Johnson JN. 1981 Dynamic fracture and spallation in ductile solids. *J. Appl. Phys.* **52**, 2812–2825. (doi:10.1063/1.329011)
19. Ortiz M, Molinari A. 1992 Effect of strain hardening and rate sensitivity on the dynamic growth of a void in a plastic material. *J. Appl. Mech.* **59**, 48–53. (doi:10.1115/1.2899463)
20. Tong W, Ravichandran G. 1995 Inertial effects on void growth in porous viscoplastic materials. *J. Appl. Mech.* **62**, 633–639. (doi:10.1115/1.2895993)
21. Molinari A, Mercier S. 2001 Micromechanical modelling of porous materials under dynamic loading. *J. Mech. Phys. Solids* **49**, 1497–1516. (doi:10.1016/S0022-5096(01)00003-5)
22. Czarnota C, Jacques N, Mercier S, Molinari A. 2008 Modelling of dynamic ductile fracture and application to the simulation of plate impact tests on tantalum. *J. Mech. Phys. Solids* **56**, 1624–1650. (doi:10.1016/j.jmps.2007.07.017)
23. Fensin SJ, Brandl C, Cerreta EK, Gray GT, Germann TC, Valone SM. 2013 Nanoscale plasticity at grain boundaries in face-centered cubic copper under shock loading. *JOM* **65**, 410–418. (doi:10.1007/s11837-012-0546-3)
24. Runnels B, Beyerlein IJ, Conti S, Ortiz M. 2016 An analytical model of interfacial energy based on a lattice-matching interatomic energy. *J. Mech. Phys. Solids* **89**, 174–193. (doi:10.1016/j.jmps.2016.01.008)
25. Castillo E. 2005 Extreme Value and Related Models with Applications in Engineering and Science. Wiley Series in Probability and Statistics. Wiley-Interscience, Hoboken, NJ.
26. Haan L, Ferreira A. 2006 *Extreme value theory: an introduction*, vol. 3. Springer. New York, NY, USA.
27. Zhang Y, Chen N, Bronkhorst CA, Cho H, Argus R. 2023 Data-driven statistical reduced-order modeling and quantification of polycrystal mechanics leading to porosity-based ductile damage. *J. Mech. Phys. Solids* **179**, 105386. (doi:10.1016/j.jmps.2023.105386)
28. AlMomani AA, Bollt E. 2020 ERFit: entropic regression fit MATLAB package, for data-driven system identification of underlying dynamic equations. (https://arxiv.org/abs/2010.02411)
29. AlMomani AAR, Sun J, Bollt E. 2020 How entropic regression beats the outliers problem in nonlinear system identification. *Chaos* **30**, 013107. (doi:10.1063/1.5133386)
30. Kim P, Rogers J, Sun J, Bollt E. 2017 Causation entropy identifies sparsity structure for parameter estimation of dynamic systems. *J. Comput. Nonlinear Dyn.* **12**, 011008. (doi:10.1115/1.4034126)
31. Elinger J. 2020 Information theoretic causality measures for parameter estimation and system identification. PhD thesis, Georgia Institute of Technology.
32. Elinger J, Rogers J. 2021 Causation entropy method for covariate selection in dynamic models. In *2021 American Control Conf. (ACC)*, pp. 2842–2847. IEEE. New Orleans, LA, USA.

33. Reynolds DA. 2009 Gaussian mixture models. *Encycl. Biom.* **741**, pp. 659–663.
34. Chakravarti IM, Laha RG, Roy J. 1967 *Handbook of methods of applied statistics*, **1**. New York: Wiley
35. Massey Jr FJ. 1951 The Kolmogorov–Smirnov test for goodness of fit. *J. Am. Stat. Assoc.* **46**, 68–78. (doi:10.1080/01621459.1951.10500769)
36. Daniel WW. 1990 Applied Nonparametric Statistics, 2nd edn. PWS-Kent Publishing Company, Boston, MA, USA.
37. Stephens MA. 1974 EDF statistics for goodness of fit and some comparisons. *J. Am. Stat. Assoc.* **69**, 730–737. (doi:10.1080/01621459.1974.10480196)
38. Aldrich J. 1995 Correlations genuine and spurious in Pearson and Yule. *Stat. Sci.* **10**, 364–376. (doi:10.1214/ss/1177009870)
39. Rohlfing I, Schneider CQ. 2018 A unifying framework for causal analysis in set-theoretic multimethod research. *Sociol. Methods Res.* **47**, 37–63. (doi:10.1177/0049124115626170)
40. Majda AJ, Chen N. 2018 Model error, information barriers, state estimation and prediction in complex multiscale systems. *Entropy* **20**, 644. (doi:10.3390/e20090644)
41. Tippett MK, Kleeman R, Tang Y. 2004 Measuring the potential utility of seasonal climate predictions. *Geophys. Res. Lett.* **31**(22). (doi:10.1029/2004GL021575)
42. Kleeman R. 2011 Information theory and dynamical system predictability. *Entropy* **13**, 612–649. (doi:10.3390/e13030612)
43. Branicki M, Majda AJ. 2012 Quantifying uncertainty for predictions with model error in non-Gaussian systems with intermittency. *Nonlinearity* **25**, 2543. (doi:10.1088/0951-7715/25/9/2543)
44. Chen N, Zhang Y. 2023 A causality-based learning approach for discovering the underlying dynamics of complex systems from partial observations with stochastic parameterization. *Physica D* **449**, 133743. (doi:10.1016/j.physd.2023.133743)
45. McLachlan GJ, Lee SX, Rathnayake SI. 2019 Finite mixture models. *Annu. Rev. Stat. Appl.* **6**, 355–378. (doi:10.1146/annurev-statistics-031017-100325)
46. Scarrott C, MacDonald A. 2012 A review of extreme value threshold estimation and uncertainty quantification. *Revstat-Stat. J.* **10**, 33–60.
47. Naveau P, Hannart A, Ribes A. 2020 Statistical methods for extreme event attribution in climate science. *Annu. Rev. Stat. Appl.* **7**, 89–110. (doi:10.1146/annurev-statistics-031219-041314)
48. Burnham KP, Anderson DR. 2004 Multimodel inference: understanding AIC and BIC in model selection. *Sociol. Methods Res.* **33**, 261–304. (doi:10.1177/0049124104268644)
49. McLachlan GJ, Rathnayake S. 2014 On the number of components in a Gaussian mixture model. *Wiley Interdiscip. Rev.: Data Min. Knowl. Discov.* **4**, 341–355. (doi:10.1002/widm.1135)
50. Figueiredo MAT, Jain AK. 2002 Unsupervised learning of finite mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**, 381–396. (doi:10.1109/34.990138)
51. Shinozaki T, Furui S, Kawahara T. 2010 Gaussian mixture optimization based on efficient cross-validation. *IEEE J. Sel. Top. Signal Process.* **4**, 540–547. (doi:10.1109/JSTSP.2010.2048235)
52. Vila JP, Schniter P. 2013 Expectation-maximization Gaussian-mixture approximate message passing. *IEEE Trans. Signal Process.* **61**, 4658–4672. (doi:10.1109/TSP.2013.2272287)
53. Chen N. 2023 *Stochastic methods for modeling and predicting complex dynamical systems: uncertainty quantification, state estimation, and reduced-order models*. Springer International Publishing, Germany.
54. Meissonnier FT, Busso EP, O'Dowd NP. 2001 Finite element implementation of a generalised non-local rate-dependent crystallographic formulation for finite strains. *Int. J. Plast.* **17**, 601–640. (doi:10.1016/S0749-6419(00)00064-4)
55. Busso EP, Meissonnier FT, O'Dowd NP. 2000 Gradient-dependent deformation of two-phase single crystals. *J. Mech. Phys. Solids* **48**, 2333–2361. (doi:10.1016/S0022-5096(00)00006-5)
56. Evers LP, Brekelmans WAM, Geers MGD. 2004 Non-local crystal plasticity model with intrinsic SSD and GND effects. *J. Mech. Phys. Solids* **52**, 2379–2401. (doi:10.1016/j.jmps.2004.03.007)
57. Counts WA, Braginsky MV, Battaile CC, Holm EA. 2008 Predicting the Hall–Petch effect in FCC metals using non-local crystal plasticity. *Int. J. Plast.* **24**, 1243–1263. (doi:10.1016/j.ijplas.2007.09.008)
58. Gao H, Huang Y. 2003 Geometrically necessary dislocation and size-dependent plasticity. *Scr. Mater.* **48**, 113–118. (doi:10.1016/S1359-6462(02)00329-9)

59. Lee S *et al.* 2023 Deformation, dislocation evolution and the non-Schmid effect in body-centered-cubic single-and polycrystal tantalum. *Int. J. Plast.* **163**, 103529. (doi:10.1016/j.ijplas.2023.103529)
60. Cho H, Bronkhorst CA, Mourad HM, Mayeur JR, Luscher D. 2018 Anomalous plasticity of body-centered-cubic crystals with non-Schmid effect. *Int. J. Solids Struct.* **139**, 138–149. (doi:10.1016/j.ijsolstr.2018.01.029)
61. Sutton AP, Balluffi RW. 1994 *Interfaces in crystalline materials*. Monographs on the Physice and Chemistry of Materials, pp. 414–423. Oxford.
62. Knezevic M, Drach B, Ardeljan M, Beyerlein IJ. 2014 Three dimensional predictions of grain scale plasticity and grain boundaries using crystal plasticity finite element models. *Comput. Methods Appl. Mech. Eng.* **277**, 239–259. (doi:10.1016/j.cma.2014.05.003)
63. Alleman C, Luscher D, Bronkhorst C, Ghosh S. 2015 Distribution-enhanced homogenization framework and model for heterogeneous elasto-plastic problems. *J. Mech. Phys. Solids* **85**, 176–202. (doi:10.1016/j.jmps.2015.09.012)
64. Cereceda D, Diehl M, Roters F, Raabe D, Perlado JM, Marian J. 2016 Unraveling the temperature dependence of the yield strength in single-crystal tungsten using atomistically-informed crystal plasticity calculations. *Int. J. Plast.* **78**, 242–265. (doi:10.1016/j.ijplas.2015.09.002)
65. Gröger R, Bailey A, Vitek V. 2008 Multiscale modeling of plastic deformation of molybdenum and tungsten: I. Atomistic studies of the core structure and glide of 1/2 ⟨111⟩ screw dislocations at 0 K. *Acta Mater.* **56**, 5401–5411. (doi:10.1016/j.actamat.2008.07.018)
66. Dezerald L, Ventelon L, Clouet E, Denoual C, Rodney D, Willaime F. 2014 *Ab initio* modeling of the two-dimensional energy landscape of screw dislocations in bcc transition metals. *Phys. Rev. B* **89**, 024104. (doi:10.1103/PhysRevB.89.024104)
67. Frederiksen SL, Jacobsen KW. 2003 Density functional theory studies of screw dislocation core structures in bcc metals. *Phil. Mag.* **83**, 365–375. (doi:10.1080/0141861021000034568)
68. Gröger R, Bailey A, Vitek V. 2008 Multiscale modeling of plastic deformation of molybdenum and tungsten: I. Atomistic studies of the core structure and glide of 1/2 ⟨111⟩ screw dislocations at 0 K. *Acta Mater.* **56**, 5401–5411. (doi:10.1016/j.actamat.2008.07.018)
69. Vitek V, Mrovec M, Bassani J. 2004 Influence of non-glide stresses on plastic flow: from atomistic to continuum modeling. *Mater. Sci. Eng. A* **365**, 31–37. (doi:10.1016/j.msea.2003.09.004)
70. Zhang Y, Dunham SD, Bronkhorst CA, Chen N. 2025 Data from: Physics-assisted data-driven stochastic reduced-order models for attribution of heterogeneous stress distributions in low-grain polycrystals. Figshare. (doi:10.6084/m9.figshare.c.7679513)