Noname manuscript No.
(will be inserted by the editor)

Process-structure-property modeling for severe plastic deformation processes using orientation imaging microscopy and data-driven techniques

Patxi Fernandez-Zelaia · Shreyes N. Melkote

Received: date / Accepted: date

Abstract Machining is a severe plastic deformation process wherein the workpiece material is subjected to high deformation rates and temperatures. During metal machining the dynamic recrystallization mechanism causes grain refinement into the sub-micron range. In this study we investigate the microstructure evolution of OFHC copper subject to a machining process where the cutting speed and rake angle are controlled to manipulate the process strain, strain rate, and temperatures. Microstructures of the deformed chips are quantified using orientation imaging microscopy and novel statistical descriptors that capture the morphology and local lattice misorientations generated during the several mechanistic stages of the dynamic recrystallization process. Mechanical properties of the resulting chips are quantified using spherical nanoindentation protocols. A multiple output Gaussian Process regression model is used to simultaneously model the structure-property evolution, which differs from more common approaches that establish such relationships sequentially. This modeling strategy is particularly attractive since it can flexibly provide both structure and property uncertainty estimates. In addition, the statistical modeling framework allows for the inclusion of multi-fidelity data. The statistical metrics utilized serve as efficient microstructure descriptors, which retain the physics of the observed structures without having to introduce ad-hoc microstructure feature definitions.

Patxi Fernandez-Zelaia
Georgia Institute of Technology
813 Ferst Dr NW
Atlanta, GA 30332
E-mail: pfz3@gatech.edu

Shreyes N. Melkote
Georgia Institute of Technology
813 Ferst Dr NW
Atlanta, GA 30332
E-mail: shreyes.melkote@me.gatech.edu

Keywords severe plastic deformation · dynamic recrystallization · materials informatics · statistics · orientation imaging microscopy

1 Introduction

Machining is a high rate severe plastic deformation (SPD) manufacturing process. The process can be described using the idealized model shown in Figure 1A. The imposed thermomechanical loading is fairly extreme with imposed strains as large as $\gamma \sim 10$, deformation rates up to $10^5\, s^{-1}$, and cutting temperatures as high as $0.6\,\theta$ (homologous temperature) [49]. These imposed deformation conditions result in microstructure refinement in both the deformed chip and the component surface [47,8,4,60,61,61,23,35]. The corresponding mechanical properties of both the chip and the workpiece surface are naturally sensitive to the produced structures [37,8,55,35]. Therefore, identifying the process-structure-property (PSP) relationships that characterize machining is critical for establishing a synergistic framework where designers, materials scientists, and manufacturers can cooperate to engineer functional surfaces. Furthermore, the SPD structures produced in machining bear a resemblance to structures produced in other SPD processes such as equal channel angular extrusion [59], high pressure torsion [65], and dynamic processes where shear banding may occur [36,34,33]. Therefore, the merit in studying machining as a high rate SPD process translates to other fields as well.

The predominant microstructure evolution mechanism in machining under ambient conditions is either continuous or discontinuous dynamic recrystallization (CDRX or DDRX) [8,37]. CDRX is driven by the formation of dislocation cells that transform to low angle boundary (LAB) sub-grain structures, and finally relative sub-grain rotations generate high angle boundary (HAB) refined grains [37,18]. DDRX is more closely related to classic recrystallization where new grains nucleate and grow, often near existing grain boundaries [26]. Since the mechanism driving CDRX is driven by lattice rotations, the structure evolution can be quantified by considering measures of crystallographic misorientation [57,1,35,51,5]. Mechanical constitutive property measurements are usually limited to hardness since the produced samples are small in scale (machined chips and workpiece surface) [37,8,23,35,55].

Materials Informatics (MI) is an emerging field within the materials community which, like cheminformatics and bioinformatics, seeks to employ statistics for addressing important domain science problems [30,29,28,38]. Materials research is conducted utilizing statistical approaches for establishing data-driven models, quantifying uncertainty, and the design and planning of experiments. MI addresses the fundamental challenge in materials research, identifying PSP relationships, by building mathematically rigorous models. The models, which may be data-driven or mixed data/physics models, may then be exploited for the design of functional materials. Recent works have established reduced-order structure-property (SP) models for single phase polycrystalline systems [43,44]. These authors utilized generalized spherical harmonics

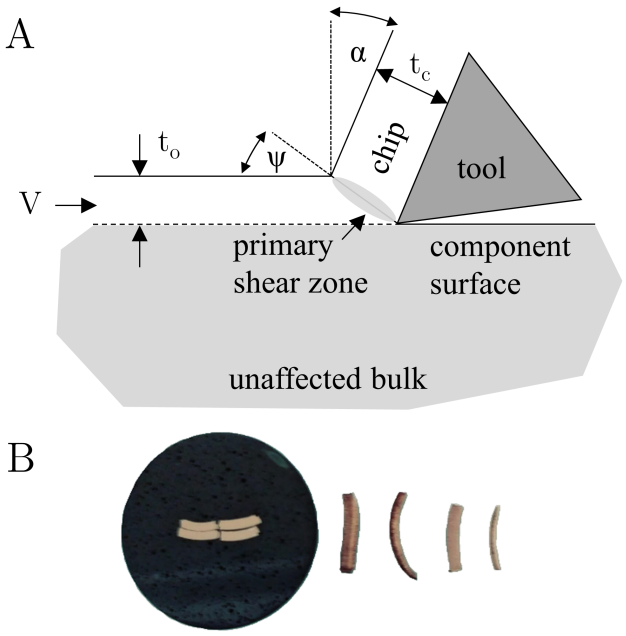


Fig. 1: (A) Machining process schematic. Controllable parameters include cutting speed (V), the uncut chip thickness (t_o), and rake angle (α). (B) Chips and metallographic sample.

(GSH) to quantify bulk textures and used spatial statistics to quantify the spatial structure describing various simulated microstructure realizations. Another recent work utilized a deep adversarial learning model coupled with a Gaussian process (GP) Bayesian design criteria for computational materials design [64].

In this work we study the evolution of pure copper subject to a high rate SPD machining process. Microstructure is quantified using orientation imaging microscopy (OIM). A microstructure statistic which quantifies the local crystal spatial misorientation is derived. This is done by utilizing a GSH basis to describe the crystallographic orientation and a unique spatial autocorrelation function, which exploits the orthogonality of the GSH basis. Constitutive mechanical properties are quantified using spherical nanoindentation tests. Finally, a Multiple Output Gaussian Process Regression (MOGPR) model is developed, which captures the full PSP relationships as well as their associated uncertainties. The model is flexible and is well suited for handling multiple *kinds* of data e.g. multi-fidelity modeling.

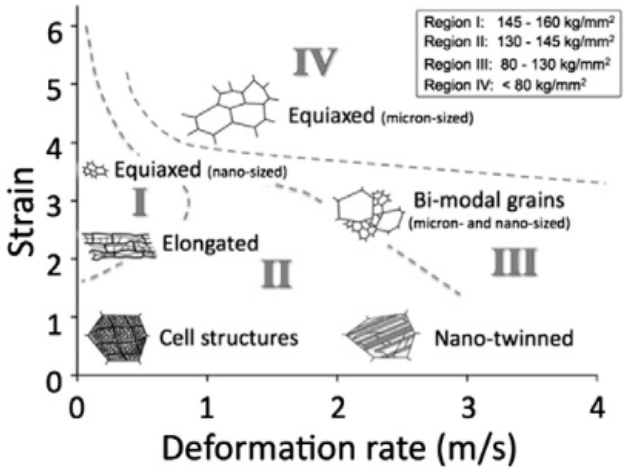


Fig. 2: Machining process-structure-property map [23].

2 Experimental Methods

Oxygen-free high conductivity copper (OFHC Cu) bars were obtained from a supplier (McMaster Carr). The material was subjected to SPD via a machining process. Tube turning experiments were carried out to emulate the idealized two-dimensional orthogonal cutting experiment shown in Figure 1A. High speed steel cutting tools with nominal rake angles $\alpha = 5°, 15°, 25°, 45°$ were used for all experiments. A constant feed (or undeformed chip thickness) (t_o) of 300 μm was prescribed for all tests. The prescribed geometry was chosen to impose large shear strains in the primary shear zone, which the machining theory predicts to be $\gamma \sim 1 - 8$ [49]. Four cutting speeds $V = 0.20, 0.33, 0.50, 1.00\, m \cdot s^{-1}$ were studied, which generate strain rates $\sim 10^3 - 10^4\ s^{-1}$. Higher cutting speeds correspondingly yield increases in chip temperatures as there is less time available for diffusion of heat away from the chip. From the measured cutting forces, the chip temperatures were estimated to reach $\sim 165°C$ for the lowest rake angle (highest strain) and fastest cutting speeds utilized [49]. Generated chips fell into a quench tank filled with water to freeze the as-machined microstructure.

Collected chips were mounted in epoxy as shown in Figure 1B. Small sample-to-sample deviations in the chip orientation within the casting will affect the perceived two dimensional morphology of observed micrographs. Furthermore, OIM results will be affected due to uncertainty in the reference sample orientation. Therefore, special care was taken to mount samples such that the observed cross section correspond as closely as possible to the idealized two dimensional orthogonal configuration. Grinding of the metallographic samples was performed to reach the "mid chip" (1 mm) thickness which is far from free boundaries and therefore minimally affected by side flow transverse

to the direction of chip flow. Samples were subsequently mechanically polished with up to $1\,\mu m$ diamond suspension polish. Final surface preparation was performed via vibratory polishing in a Buehler VibroMet 2. A Tescan Mira XMH field emission scanning electron microscope (FE-SEM) was utilized to image the generated microstructures. A backscatter emissions (BSE) detector was utilized for all imaging as it was found to yield images with extremely good contrast (see Figure 3). A EDAX Hikari EBSD detector with TSL OIM analysis was utilized for orientation imaging.

Nanoindentation experiments were performed on a Agilent G200 nanoindenter with an XP head and continuous stiffness monitoring (CMS). A $100\,\mu m$ diamond indenter was used for all experiments. Spherical indentation stress-strain protocols were utilized to further process experimental data [41,42]. The derived indentation stress-strain curves capture the mechanical response of the material deformed beneath the indenter. The corresponding contact radius for these experiments varied between $10 - 20\,\mu m$. The microstructures considered vary greatly in their degree of refinement. Under some conditions, very fine structures ($d < 1\,\mu m$) were generated suggesting that the obtained indentation responses are likely well homogenized. Coarser structures however suggest that the local material heterogeneity may introduce additional response variation. In our analysis we will account for this by attempting to establish *mean* property quantities.

Microhardness measurements were performed using a Buehler series 1600 microhardness tester. A diamond tip Vickers indenter loaded to $500g$ was used for all tests.

3 Methods

3.1 Microstructure quantification

BSE and EBSD micrographs for two different process conditions are shown in Figure 3. Images at larger values of the rake angle α (or smaller values of strain since $\gamma \propto \alpha^{-1}$) produced correspondingly coarser microstructures and therefore larger fields of view were required at these settings. The field of view at each setting is illustrated in Figure 4. The total number of raster steps in each image was maintained at 300×300 to avoid unnecessarily long scans.

In Figure 3, it is clear from both the BSE and EBSD scans that the microstructures are morphologically different. In the $\alpha = 25°$ BSE image however it is difficult to discern which features are grain boundaries; the BSE image is sensitive to defect structures besides grain boundaries. An even clearer pattern is visible in Figure 4 particularly at low rake angles of 5° and 15°; with increasing cutting speed it appears as if the structure becomes *smeared*. Statistically, it can be stated that crystal orientations are more *spatially correlated* at higher cutting speeds than at lower cutting speeds. Furthermore, this pattern is also present with increasing rake angle. Consider an experiment where a point is chosen randomly in the micrograph for (5°, $1.00m \cdot s^{-1}$) and we

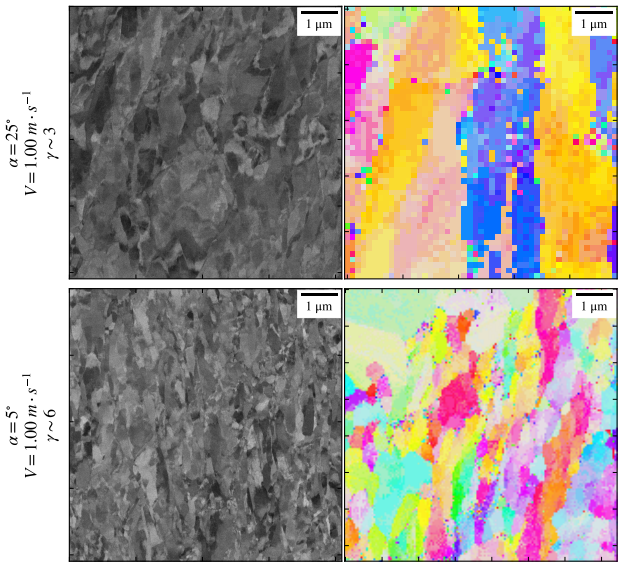


Fig. 3: BSE-SEM and EBSD images of the generated microstructures. Top images correspond to process conditions that impose less strain relative to the bottom images. BSE and EBSD images are not coincident.

note the crystal orientation at the chosen pixel and at a location 5 μm to its right. Subsequently, the same experiment is performed on the micrograph for (5°, $0.20m \cdot s^{-1}$). On average, over many repetitions, the two pixels from (5°, $1.00m \cdot s^{-1}$) would yield more "similar" orientations than in the micrograph for (5°, $0.20m \cdot s^{-1}$). It is this feature that we wish to quantify and exploit for assessing microstructural anisotropy.

Recent advances in the MI community have established statistically rigorous methods for quantifying stochastic material systems [28]. In this work we quantify microstructure via crystallographic orientation which can be quantified using the Bunge-Euler angles $\boldsymbol{g} = (\phi_1, \Phi, \phi_2)$, which are continuously defined over the fundamental zone (FZ) [9]. The probability of finding orientation $\boldsymbol{g}$ at spatial location $\boldsymbol{x}$ is $f_x(\boldsymbol{g})$ [63,43]. Note that in the MI literature this quantity is referred to as the microstructure function [2].

Spatial correlations between microstructure states can be quantified through the use of spatial statistics [56,28]. The simplest of the n-point spatial statistics is two-point statistics. These quantities capture spatial correlations by considering the vector distance between two points. The example posed earlier in this section used two-point statistics to qualitatively describe the "spread" of crystals. Formally, the two-point statistics can be described by a conditional probability,

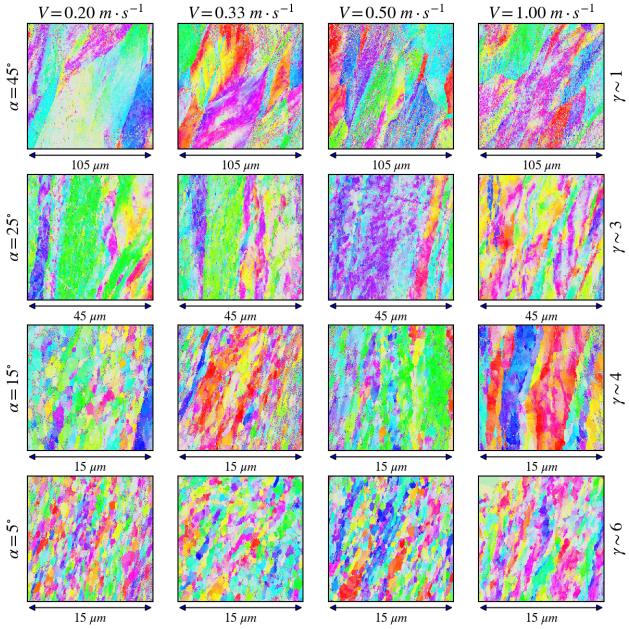


Fig. 4: EBSD images of the various microstructures produced via machining.

$$p\left(\boldsymbol{g}, \boldsymbol{g}'|\boldsymbol{t}\right) = \frac{1}{|\mathcal{X}|} \int_{\mathcal{X}} f_{\boldsymbol{x}}\left(\boldsymbol{g}\right) f_{\boldsymbol{x}+\boldsymbol{t}}\left(\boldsymbol{g}'\right) d\mathcal{X}, \tag{1}$$

where $\boldsymbol{t}$ is the vector that separates two points in the microstructure, $\boldsymbol{g}$ is the microstructure state at the tail of the vector, and $\boldsymbol{g}'$ is the microstructure state at the head of the vector. Note that if $\boldsymbol{g}' = \boldsymbol{g}$ then this quantity describes *autocorrelation*. Also note that this quantity is solely a function of the *difference* in spatial location between two points ($\boldsymbol{t}$) and therefore this definition assumes *stationarity* of the microstructure.

Consider now that we wish to obtain a compact representation of $f_{\boldsymbol{x}}(\boldsymbol{g})$. There have been several works that have adopted the use of generalized spherical harmonics (GSH) for describing this quantity in polycrystalline systems [63,43,44]. Using a GSH basis $f_{\boldsymbol{x}}(\boldsymbol{g})$ can be rewritten as,

$$f_{\boldsymbol{x}}\left(\boldsymbol{g}\right) = \sum_{\mu,n,l} F_{l\boldsymbol{x}}^{\mu n} \dot{T}_{l}^{\mu n}\left(\boldsymbol{g}\right), \tag{2}$$

where μ, n, l represent multiple indices for multiple sums, and $F_{l\boldsymbol{x}}^{\mu n}$ is the complex-valued GSH coefficient at $\boldsymbol{x}$ which corresponds to the complex valued GSH basis $\dot{T}_{l}^{\mu n}$. Note that the $\dot{T}_{l}^{\mu n}$ preserve crystal symmetries and are orthogonal to their complex conjugate $\dot{T}_{l}^{\mu n*}$. The coefficients $F_{l\boldsymbol{x}}^{\mu n}$ can be obtained in the analogous way to how Fourier coefficients are determined (i.e.

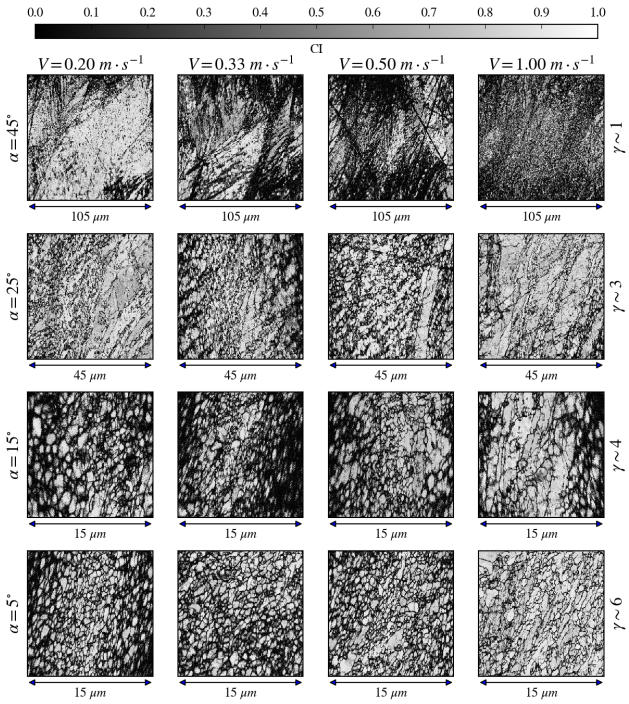


Fig. 5: Confidence index maps corresponding to each EBSD scan.

by exploiting orthogonality) $F_{l\boldsymbol{x}}^{\mu n} = (2l+1)\int f_{\boldsymbol{x}}(\boldsymbol{g})\,\dot{\bar{T}}_l^{\mu n*}(\boldsymbol{g})\,d\boldsymbol{g}$. In the case where spatial bin $\boldsymbol{x}$ is occupied by a single orientation $\boldsymbol{g}_o$ (an individual pixel in an indexed EBSD scan) then $F_{l\boldsymbol{x}}^{\mu n} = (2l+1)\dot{\bar{T}}_l^{\mu n*}(\boldsymbol{g}_o)$.

Naturally, the next step is to redefine the two-point statistics using the GSH basis representation. One practical consideration is that there are an infinite number of $\boldsymbol{g}$ to chose from since it is a continuous function. In recent works, this problem is overcome by computing spatial statistics over the complex valued GSH coefficients themselves [43,44]. The interpretation is that the different microstructure states are described by the different GSH coefficients indexed over μ, n, l. However, in this work we will introduce one additional definition which produces a different interpretation of the spatial statistics. Here we define an averaged quantity for the spatial autocorrelation, which averages over all $\boldsymbol{g}$. In doing so, information about texture is lost, but this new definition is well suited for capturing the local *misorientation* or local *morphological* spatial behavior. Therefore, we define,

$$\bar{p}_{\boldsymbol{t}} = \frac{1}{V_{FZ}}\int_{FZ} p(\boldsymbol{g},\boldsymbol{g}|\boldsymbol{t})\,d\boldsymbol{g}, \tag{3}$$

where V_{FZ} is the fundamental zone volume. Again, this quantity describes the spatial autocorrelation of crystallographic orientation *averaged* over all possible crystal orientations. Some information (texture) is lost but the structural morphological information is retained. The advantage of adopting this strategy is that often very large scans are needed to capture texture, which is inherently a volume-averaged quantity. Therefore, texture requires a large representative volume element (RVE) to be statistically representative of the material as a whole. Conversely, local morphological features may be representative at much smaller RVE length scales.

Combining the GSH representation of Eqn. 2, definition in Eqn. 3, and two point statistics in Eqn. 1 the following expressions may be derived:

$$\begin{aligned}
\bar{p}_{\boldsymbol{t}} &= \frac{1}{|\mathcal{X}|}\frac{1}{V_{FZ}}\int_{\mathcal{X}}\int_{FZ} f_{\boldsymbol{x}}\left(\boldsymbol{g}\right) f^{*}_{\boldsymbol{x}+\boldsymbol{t}}\left(\boldsymbol{g}\right) d\boldsymbol{g}d\mathcal{X} \\
&= \frac{1}{|\mathcal{X}|}\frac{1}{V_{FZ}}\int_{\mathcal{X}}\int_{FZ} \sum_{\mu,n,l} F^{\mu n}_{l\boldsymbol{x}}\dot{T}^{\mu n}_{l}\left(\boldsymbol{g}\right) \sum_{\mu,n,l} F^{\mu n*}_{l\boldsymbol{x}+\boldsymbol{t}}\dot{T}^{\mu n*}_{l}\left(\boldsymbol{g}\right) d\boldsymbol{g}d\mathcal{X} \\
&= \frac{1}{|\mathcal{X}|}\frac{1}{V_{FZ}}\frac{1}{2l+1}\int_{\mathcal{X}} \sum_{\mu,n,l} F^{\mu n}_{l\boldsymbol{x}} F^{\mu n*}_{l\boldsymbol{x}+\boldsymbol{t}} d\mathcal{X} \\
&= \frac{1}{|\boldsymbol{S}|}\frac{1}{V_{FZ}}\frac{1}{2l+1}\sum_{\boldsymbol{s}=1}^{\boldsymbol{S}}\sum_{\mu,n,l} F^{\mu n}_{l\boldsymbol{s}} F^{\mu n*}_{l\boldsymbol{s}+\boldsymbol{t}}
\end{aligned} \tag{4}$$

where $f^{*}_{\boldsymbol{x}+\boldsymbol{t}}\left(\boldsymbol{g}\right)$ is the complex conjugate. Since f is a real valued function then $f = f^{*}$. This trick enables significant simplification when computing the product of the two large sums since we are using an orthogonal basis; $\int_{FZ}\dot{T}^{\mu n}_{l}\dot{T}^{\mu' n'*}_{l'}d\boldsymbol{g} = (2l+1)^{-1}$ if all the indices "match" else 0. A similar manipulation was found in [63] but in their case it was for computing localization relationships and not spatial autocorrelations. In fact, the definition introduced in Eqn. 3 was purposefully introduced to exploit the orthogonality found in the GSH basis similar to what was done in [63]. This simplification only works for the case of autocorrelation; the orthogonality cannot be exploited when considering cross correlations. The final expression obtained is a function (mean) of the autocorrelation statistics derived in [43,44]. However, our derivation can be justified with some novel physical interpretation (mean autocorrelation over all $\boldsymbol{g}$).

Note that although $f_{\boldsymbol{x}}(\boldsymbol{g})$ is described using a truncated GSH expansion each of the GSH coefficients themselves is a complex-valued continuous variable. This treatment allows for gradations of similarity between pixels. For instance, pixels misoriented by only a few degrees will yield higher autocorrelation than pixels with large misorientation. If instead the continuous-valued microstructure state (orientation $\boldsymbol{g}$) was discritized using a "binning" strategy [17], then pixels with similar orientations that happen to fall into different bins would erroneously suggest a lack of autocorrelation. Furthermore, binning of

the three dimensional orientation space would be cumbersome and inefficient [63].

The final line of Eqn. 4 discretizes the spatial domain over $\mathcal{X}$ into a two-dimensional binned spatial domain over $\mathcal{S}$ which corresponds to the EBSD scan pixels. The final expression is a convolution over $\boldsymbol{S}$, which can be efficiently computed using discrete Fourier transforms (DFTs) [28]. The quantity $|\boldsymbol{S}|$ is the total number of spatial bins considered e.g. total number of pixels in a image. Note for partial scans, scans where a portion of the image contains unreliable or "bad" measurements, recent algorithms have been established that account for this complication by modification of Eqn. 4 [11].

The proposed microstructure descriptor is sensitive to the degree of GSH discretization introduced in Eqn. 2. If too few terms are used in the sum then it may be possible that $f_{\boldsymbol{x}}(\boldsymbol{g})$ will be unable to accurately describe certain orientations present in the observed micrographs. Consequently, the morphologies associated with those inadequately resolved orientations will be neglected in the mean autocorrelation function (Eqn. 4). There are two recent works which address the question of GSH truncation when quantifying spatial microstructure data. Paulson et al. published a work on the homogenization of elastic and inelastic properties of polycrystalline HCP systems using a similar MI approach [43]. For HCP systems $l = (0, 2)$ yields 6 terms and $l = (0, 2, 4)$ yields 15 terms in Eqn. 2. In their study they considered various crystallographic textures and found that truncation at 15 terms yielded marginally better results than at 6 terms. Yabansu, Patel, and Kalidindi found that truncation with $l = (0, 4)$, yielding a total of 10 terms, was suitable for building reduced order elastic localization relationships in polycrystalline FCC systems [63]. Therefore, since there is evidence that both localization and homogenization relationships can be captured with minimal terms, we argue that a ten term GSH truncation should be sufficient for adequately describing the microstructures studied in this work.

3.2 Feature selection and bootstrapping

The previous section describes a rigorous method for quantifying the microstructure. The mean two point statistics, $\bar{p}_{\boldsymbol{t}}$, derived however is of the same dimensionality as $\boldsymbol{t}$. Correspondingly, $\boldsymbol{t}$ is a vector that can be placed into the microstructure and hence in this case it is bounded by the size of the EBSD scans/images. Therefore, $\bar{p}_{\boldsymbol{t}} \in \mathcal{R}^{N \times M}$ where N and M are the height and width of the images measured in pixels. All EBSD scans in this work are square hence the dimensionality of each statistic derived from the images is N^2. Therefore, it is clear that for interpretability of the results some dimensionality reduction will be necessary. In this work we utilize unsupervised Principal Component Analysis (PCA), which computes a statistically optimal basis for describing the full feature space. PCA has been employed successfully in many MI works for compact representation of microstructure statistics [14, 28, 12, 27, 44, 43, 32, 53, 54]. Dimensionality reduction is achieved by suitably truncating

the basis expansion and using the basis weights (PC weights) to describe the data. This is analogous to Fourier representation of a one dimensional signal where the Fourier coefficients can compactly describe the signal.

Another consideration when constructing the microstructure feature space is the need to ensure rotational invariance of the images. Consider that small deviations in how the samples are mounted in the microscope may result in angular rotation of the images, which therefore affects the microstructure statistics. Looking ahead at Figure 9, careful inspection reveals that the statistics are slightly rotationally misoriented relative to one another. Failure to capture this experimental artifact could result in falsely discriminating two otherwise statistically identical microstructures. Rotational invariance is introduced by utilizing the methods found in [14]. Full details of this method are found in the referenced work and are not reproduced here.

Finally, a strategy is needed to obtain measurements of the *dispersion* of the PC weights. A naive and experimentally costly strategy would require that multiple EBSD scans be taken. From the dispersion (variance, covariance) measures, hypothesis testing could be performed or data-driven models could be built. This approach would be extremely expensive as each single scan is costly to obtain. An alternative strategy is to use the single observations and obtain dispersion estimates from *bootstrapping* of the images [16]. A similar strategy was utilized in [62,13] for generating computationally efficient statistical volume elements (SVEs). Niezgoda, Yabansu, and Kalidindi utilized bootstrapping to obtain estimates of the structural variance of three dimensional simulated microstructures [39].

Bootstrapping seeks to establish dispersion estimates for mean quantities by a resampling of the data [16]. It is appealing because no distributional assumptions are needed (e.g. normality). Furthermore, it can be used to obtain dispersion estimates for complicated functions of the observed data. Consider that we make N observations of a normally distributed quantity X but we want the mean and mean-variance of some complicated function $f(X)$.

In our setting, the data are the EBSD scans and the transformation is the pipeline that transforms the EBSD scans to $\bar{p}_{\boldsymbol{t}}$ and then to the truncated PC weights. Special care is also needed to preserve the spatial correlation structure present in our data. Therefore, we used a strategy that is analogous to bootstrapping time-series data [16]: $7.5\mu m \times 7.5\mu m$ images were sub-sampled four times and used to reconstruct a tiled $15\mu m \times 15\mu m$ image. The $7.5\mu m \times 7.5\mu m$ tiles were obtained by randomly selecting pixels from the image and then obtaining $3.75\mu m$ worth of pixels left, right, above, and below the selected point. In the case where the randomly selected pixel was within $3.75\mu m$ of a boundary the sub-sampled image was obtained by "wrapping" around the original image. For computing spatial statistics this is acceptable since the convolution in Eqn. 4 assumes periodic boundary conditions, which is equivalent to assuming that the image "wraps" around itself. This is shown schematically in 7. Each resampled $15\mu m \times 15\mu m$ image corresponds to a single bootstrapped sample. For each setting, 100 bootstrapped samples were generated. The entire ensemble was then utilized to establish a PC basis and the corresponding

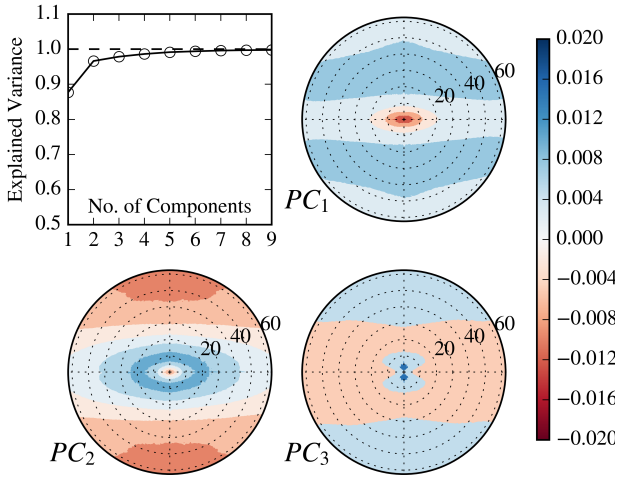


Fig. 6: Rotationally invariant mean spatial crystallographic autocorrelation basis and accumulated variance explained statistics.

PC weights for each bootstrapped sample were determined. The mean and variance of these bootstrapped PC weights were utilized to establish the mean and mean-dispersion at each unique process setting.

3.3 Multiple output Gaussian process regression

A data-driven model is needed to efficiently map the controllable process parameters to the material quantities of interest. In this setting, the structure behaves as an intermediate variable that fundamentally controls the physics and is responsible for the exhibited properties. A statistical interpretation is that the structure variable is a *latent* variable; it is critically important but is either not possible to observe or perhaps can only be observed with great effort. This is an important consideration when identifying the relevant length scales and corresponding salient microstructural features. For instance, consider that TEM micrographs are rich with information at the lowest length scales but are costly to obtain. Conversely, optical micrographs are relatively easy to obtain but may have limited utility for certain problems, for instance properties that are dependent on the lower length scale physics. Process-property models can sometimes capture the underlying relationships [15], however, inclusion of structure into the modeling pipeline is preferred [28]. The justification is that structure physically governs the underlying property behavior and inclusion of such information may alleviate potential ambiguities associated with non-unique process-property mappings.

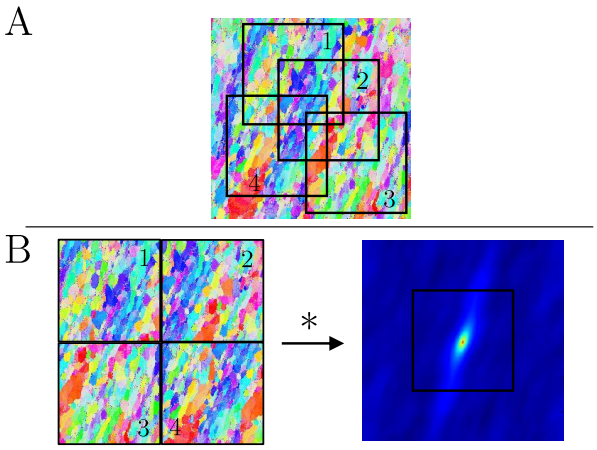


Fig. 7: Bootstrapping schematic for estimating confidence bounds on mean feature statistics. (A) original EBSD scan and corresponding random samples (B) reconstruction from random sampling and associated bootstrapped mean spatial crystallographic autocorrelation $\bar{p}_{\boldsymbol{t}}$ sample.

Modeling of PSP relationships traditionally follows a sequential strategy where the process-structure (P-S) and structure-property (S-P) relationships are established independently of one another and combined in sequence [32]. This is illustrated in the top of Figure 8. A difficultly associated with such a framework is that it is not straightforward to quantify uncertainty propagation between P-S and S-P models. The P-S model accepts process parameters as inputs, which are considered to be deterministic. The output microstructure estimates are naturally stochastic since the microstructure observations are stochastic. Computing confidence bounds for the output microstructure estimates is trivial in most statistical frameworks. Additional care however is needed in the subsequent S-P modeling step when transferring forward stochastic structure estimates. As was just argued, the P-S outputs are stochastic and hence the S-P inputs are stochastic. However, most data-driven models assume the model inputs to be deterministic.

Another limitation of the sequential PSP modeling strategy is that information is not *shared* across the P-S and S-P models. Consider that the model of interest is actually the full PSP model. This model is of course built using the two P-S and S-P sub-models, which are usually established independently. A better PSP model could perhaps be established if the P-S and S-P models were built concurrently or perhaps with iteration; the best P-S and S-P models established independently may not produce the best PSP model.

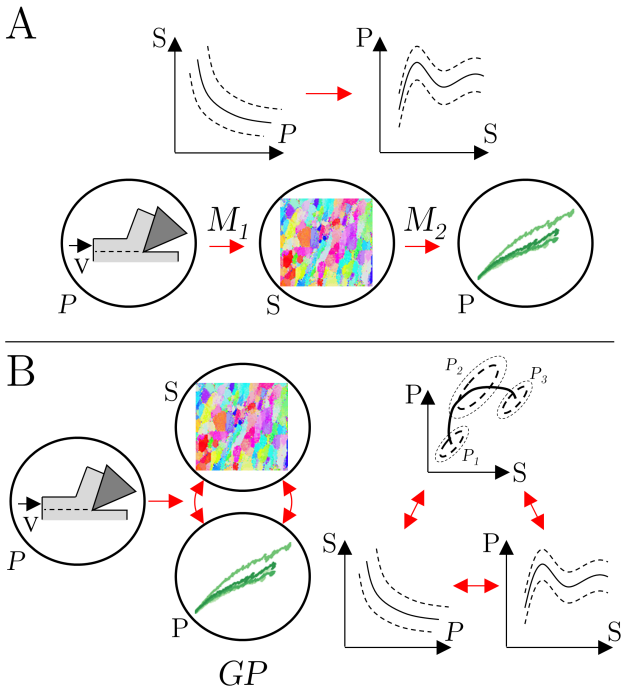


Fig. 8: Schematic of two modeling strategies for establishing PSP-linkages. Note that italicized P refers to the process and normal font P represents properties. (A) A sequential strategy where process-structure and structure-property models are built independently and predictions flow sequentially, (B) jointly developed model using multiple output GP structure, which captures possible cross-correlations in the structure-property structure.

Gaussian process regression (GPR) is a non-parametric curve fitting technique [46,48]. Unlike traditional linear and nonlinear regression, non-parametric methods do not require a priori knowledge of the trends' functional form. Instead, the data is assumed to come from a Gaussian data generating process where observations, y_i and y_j, may be correlated based on their proximity to each other, $\boldsymbol{x}_i - \boldsymbol{x}_j$. Future predictions, $y(\boldsymbol{x})$, can be shown to be a weighted average of all the $[y_1, \ldots, y_N]$ where the weights depend on the proximity of $\boldsymbol{x}$ relative to all observations in the dataset $[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]$. The final form of the GPR statistical model is closely related to kernel regression and smoothing methods [6,46].

In classical regression, the statistical inference or learning is performed by optimally estimating the unknown regression coefficients. In the GPR setting the inference is performed by estimating the unknown statistical *hyperparameters*. These quantities define the *correlation structure*, which is embedded in

the collected observations. For instance, correlation length scales are used to precisely quantify the relative measures of "proximity" mentioned above. In some problems $x_i - x_j = 1$ may be an insignificant difference yet in other instances this may be large.

In this work we attempt to address both these considerations by utilizing a multiple output Gaussian process regression (MOGPR) model for simultaneously identifying the full PSP model (Figure 8B). Multiple output implies that y need not be a scalar. This model choice offers several promising features not available using a sequential strategy. Firstly, structure and properties are modeled together as a function of process inputs using a multivariate normal structure to quantify structure-property correlations. In this way process-property is possible however the model will also infer possible structure-property correlations when present. The structure-property cross-correlation jointly considers the full PSP linkage rather than independent sub-models. Secondly, quantifying the S-P variables simultaneously in a multiple output setting allows for easy uncertainty quantification of all relevant quantities including their cross-correlation structure. Finally, the MOGPR framework is flexible in its treatment of data and enables the inclusion of partial datasets with missing data. For instance consider a study where there are two microstructure descriptors. One is obtained using efficient experiments such as optical microscopy. The other descriptor is obtained using TEM and is therefore costly to acquire. The dataset may therefore contain many times more optical images than TEM images. However, in establishing the S-P linkages standard regression models require *both* covariates for each individual property measurement. Clearly, such a framework cannot pair the two descriptors since one is much more numerous! The state of the art in this setting is to implement a *transfer learning* model which enables sharing of information between the two *kinds* of structure data [40]. The MOGPR model can automatically accommodate this setup. Additional details on the GPR framework, estimation of hyperparameters, prediction estimates, and details on the implementation used in this work are found in A.

3.4 Multi-fidelity property modeling

In this work, structural descriptors come from the PC-weights of the mean crystallographic autocorrelation function ($\hat{p}_{\boldsymbol{t}}$). Property measurements are obtained using spherical nanoindentation. The indentation stress-strain yield strength is used to quantify material strength [41]. The fraction explained variance (Figure 6) illustrates that two PC components capture 97% of the observed variance. Therefore, in this study $M = 3$ where $j = 1, 2$ are the first two PC-weights and $j = 3$ is the indentation yield e.g. the MOGPR model represents the vector (PC_1, PC_2, Y_{ind}). Y_{ind} has some physically meaningful interpretation but is a somewhat noisy observation (see Figure 17). This variation is inherited from various sources including microstructure and surface characteristics. In Figures 3 and 4 it is clear that the indenter could possibly

engage different crystallographic orientations from test to test. Furthermore, there is also morphological heterogeneity across microstructures as seen in Figure 5. Although the final contact radius using a $100\mu m$ indenter is on the order of 10-20 μm, the contact radius is at the yield point roughly 1-2μm. Even using a larger $500\mu m$ indenter would not produce RVEs of crystallographic orientation and larger indenters (the next available indenter is 1500 μm) are not feasible due to the load-limits of the machine and the size of our samples (the smallest is 500 μm in thickness). A brute-force strategy would require EBSD imaging of every SVE indentation site, which is experimentally costly. Finally, the response is sensitive to nano-scale asperities on the prepared surfaces, which introduces variation in the form of noise.

Therefore, in this work our strategy is to simply homogenize over these effects and therefore we have conducted many repeated indentation experiments for each unique process setting. However, a complimentary strategy is available that allows the combination of nanoindentation data with cheaper lower-fidelity property data. In the statistics community this is referred to as *multi-fidelity* modeling [24,58,31]. For this work we consider the Vickers microhardness (HV) as a cheap property measure. The justification is that the spherical indentation stress-strain protocols enable granular interpretation of both elastic and post-elastic behavior of the indented material whereas hardness does not. Nevertheless, microhardness is easy to obtain and therefore may aid in bolstering confidence in our inferences. Additionally, the hardness data is less noisy because it is less sensitive to the previously described heterogeneities since the volume of material probed is much larger; diagonals produced during indentation at $500g$ load were on the order of 80-100 μm. A key assumption here is that Y_{ind} and HV follow the same trends. We will introduce some flexibility, however, in case they do not follow the same trends or if they do not follow the same trends under certain process settings. The necessary statistical framework for incorporation of multi-fidelity property data may be found in B.

4 Results

The mean crystallographic autocorrelation for each micrograph is shown in Figure 9. Note that these autocorrelation statistics are *empirical* quantities as they are computed directly from the data using Eqn. 4, which is free from any parametric assumptions. It is important to acknowledge this as subsequent modeling is performed by directly comparing these statistics and therefore the same field of view (FOV) must always be used. All the statistics shown in Figure 9 have a field of view of 15 μm. Therefore, images obtained at $\alpha = 25°, 45°$, which have FOV of 45 and 105 μm, were sub-sampled. The analysis therefore does not consider autocorrelation information available at larger correlation lengths in these images. However, this "clipping" is necessary to maintain identical scales across all the empirically computed autocorrelations.

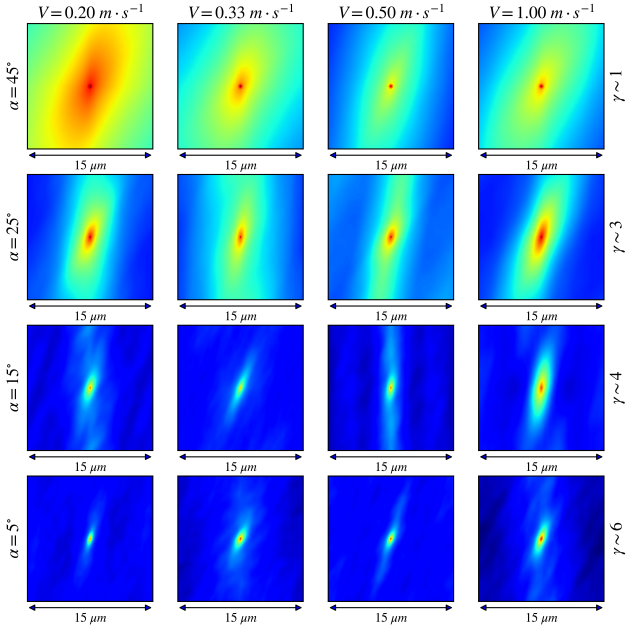


Fig. 9: Mean spatial crystallographic autocorrelation $\bar{p}_{\boldsymbol{t}}$ for each process setting. Note that for direct comparison of these statistics must be over the same length scale therefore larger image statistics cropped down to 15 μm.

Bootstrapped samples of the rotationally invariant mean crystallographic autocorrelation are shown in Figure 10. Recall that 97% of the variance can be captured with a truncated PCA expansion using only two principal components, see Figure 6. Also shown in Figure 10 is the predicted MOGPR path in PC-space. The bootstrapped samples visually appear to generate scatter close to a bivariate normal distribution. Both the degree of scatter and the correlation in the scatter varies for each unique process setting. Therefore, the components of the observation error covariance matrix, $\boldsymbol{\Sigma}$, which correspond to these structural variables were prescribed using frequentist estimates for each unique process setting. This simplification is justified since the scope of our work is to quantify and model *mean* quantities. Additionally, bootstrapping is an effective method for estimating the dispersion of statistics and therefore the hyperparameter inference in Eqn. A.8 can be simplified. Furthermore, since the repetitions themselves only capture dispersion information of the data, and the observation error is specified, it is only necessary to utilize the mean value structure variables, $\bar{\boldsymbol{PC}}_i$, when building the MOGPR model. This final point saves a great deal of computational burden associated with inverting $\boldsymbol{C} + \boldsymbol{\Sigma}$. This simplification requires only 16 two-dimensional mean values rather than the full data set.

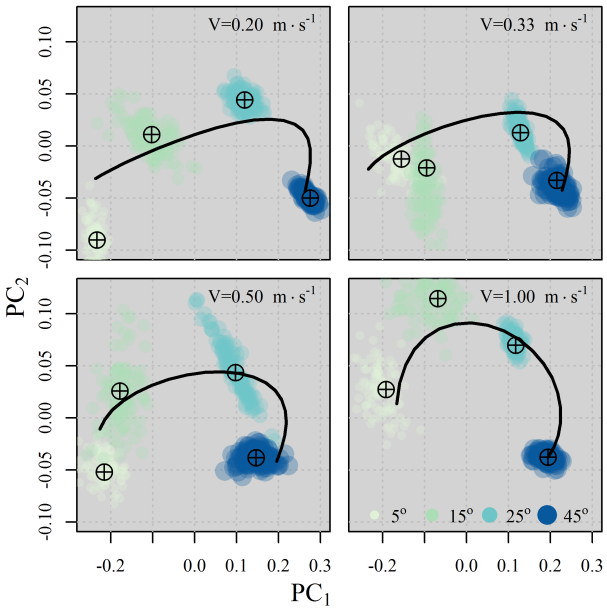


Fig. 10: Mean PC_1 and PC_2 evolution over process settings and GP model path prediction. Shown data are the 100 bootstrap samples at each process setting and each corresponding mean ($\oplus$).

In Figures 11 and 12 the structure-property relations are shown. Note that structure-property data are not paired; there is not a "corresponding" property measure for each micrograph. Visualization however requires pairing and therefore the mean values and the associated confidence intervals are shown for experimental data. The mean MOGPR path and the confidence *region* are also shown. Note that there is a clear distinction between the confidence region of the mean and confidence region of future observations. Future observations will also contain some observation errors and would therefore have a correspondingly larger confidence region. At $V = 1.00\ m \cdot s^{-1}$ the trends appear to change despite the behavior being fairly consistent across cutting speeds $V < 1.00\ m \cdot s^{-1}$. This experimental setting corresponds to the largest imposed temperatures since $\Delta t \sim 1/V$ and hence there is less time available for conduction of heat away from the generated chips [50].

Process-structure relationships are shown in Figures 13-16. It is clear that the rake angle, α, has the greatest influence on the generated structures. This agrees with intuition as α controls the geometric configuration of the experiment and therefore has the greatest impact on the imposed shear strains γ. Deformation conversely drives structural refinement and evolution via the DRX mechanism [8].

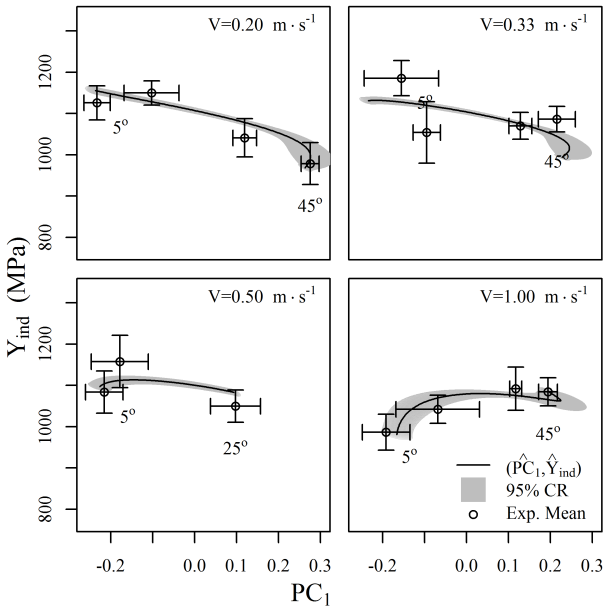


Fig. 11: Mean PC_1 and Y_{ind} evolution over process settings and GP model path prediction and 95% confidence region. Error bars correspond to mean variation for Y_{ind} and the bootstrapped variation for PC_1.

Finally, the process-property maps are shown in Figure 17. Note that process-property *implicitly* considers structural relationships via the MOGPR model. The Vickers hardness data generally follows trends similar to the indentation yield. At the highest cutting speed, $V = 1.00\ m \cdot s^{-1}$, there is a significant decrease in hardness/strength going from $\alpha = 15°$ to $\alpha = 5°$. This is only observed at the highest speed, which suggests that physically this anomalous behavior is driven by thermal effects.

5 Discussion

The proposed mean crystallographic autocorrelation spatial statistic is an effective measure of microstructural morphology. The power of this metric is that it quantifies morphology without the need to explicitly define microstructural features. A common assumption when analyzing EBSD data is to define a threshold misorientation value for defining high angle boundaries. At other times, the misorientation distribution function (ODF) itself is utilized as a metric but this necessitates identification of grain boundaries, which is again based on assumed threshold values [60]. Since our statistic only captures morphological features it may be well suited in settings where the scan size

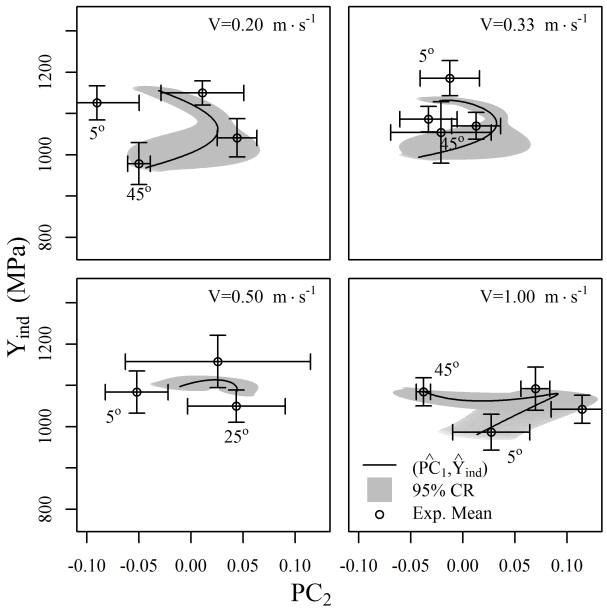


Fig. 12: Mean PC_2 and Y_{ind} evolution over process settings and GP model path prediction and 95% confidence region. Error bars correspond to mean variation for Y_{ind} and the bootstrapped variation for PC_2.

is smaller than what is required for accurately quantifying texture. Crystallographic texture is a homogenized quantity and therefore larger scans are typically necessary to accurately capture the representative crystallographic texture. The $15\,\mu m \times 15\,\mu m$ images in Figure 4 are certainly not sufficient for identifying texture but can still be used for quantifying morphological features.

Physical interpretation of the obtained microstructure evolution results is possible by considering the PCA bases shown in Figure 6. Recall that $\bar{p}_{\boldsymbol{t}}$ measures the degree of spatial crystallographic autocorrelation (similarity). The first principal basis corresponding to PC_1 is highly localized with large negative values towards the center of the basis, some positive asymmetric values away from $\theta = 0°$, and slightly positive in the remainder of the region. The peaked negative region corresponds to a length of about 10 pixels which is $500\,nm$ ($50\,nm/pixel$). Note that this corresponds to the refined crystallite size observed at the largest strains. Conversely, PC_2 has an even sharper, but faint, negative peak in the center, positive values in the $0.5 - 2\,\mu m$ range, and negative values at large distances. Therefore, one contribution of the PC_1 basis is to control a high autocorrelation region concentrated within a $500\,nm$ region. PC_2 captures competing autocorrelation trends in the $0.5 - 2\,\mu m$ and $> 3\,\mu m$ range. Therefore, it is reasonable that PC_1 is observed to displays the greatest sensitivity to the applied rake angle (Figures 13 & 14). As the rake

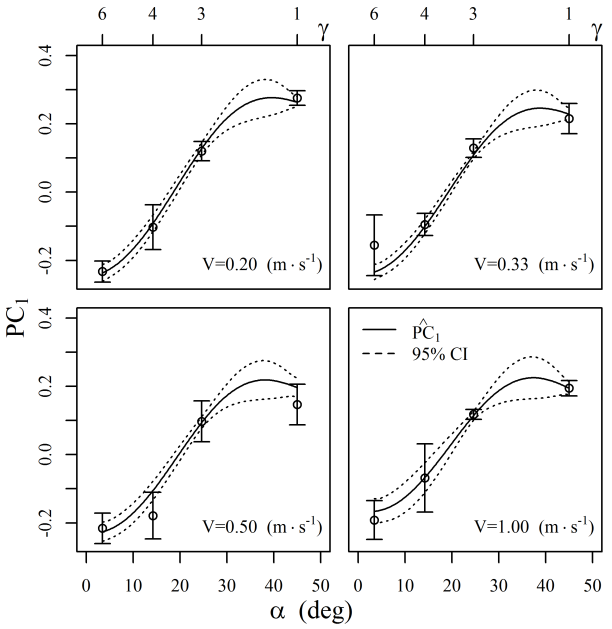


Fig. 13: Mean PC_1 evolution versus α and the corresponding GP model prediction and 95% confidence bounds. Error bars correspond to the bootstrapped variation for PC_1.

angle is decreased, strains are increased, DRX drives refinement, and therefore pixels only retain autocorrelation with very close neighbor points (roughly within a crystal). However, PC_1 does not appear to significantly change with cutting speed (Figure 15). This is because cutting speed does not influence spatial similarity at these small scales. PC_2 however does appear to be sensitive to cutting speed (Figure 16) and this sensitivity decreases with increasing rake angle (decreasing strain). This implies that at large imposed strains, as the cutting speed is increased, similarity of crystal orientation extends to include larger neighborhoods in the $0.5 - 2\,\mu m$ region. This observation agrees with the process physics where it is known that cutting temperatures increase with both increasing speeds and strains. Additional straining drives heat generation via plastic dissipation and increased cutting speeds limit the efficacy of conduction to remove heat away from the process zone. At higher temperatures DRX is less impactful [8] and thus there is less misorientation and hence crystal similarity extends over larger spatial distances (less misorientation). Therefore, PC_2 is sensitive to thermal effects, which are implicitly tied to the cutting speeds. With respect to the rake angle, PC_2 has a significant quadratic interaction and this complex behavior may be explained as follows. At high rake angles (low strains) the similarity extends over large distances ($> 3\,\mu m$) and PC_2 is negative, which yields large positive autocorrelation values at large

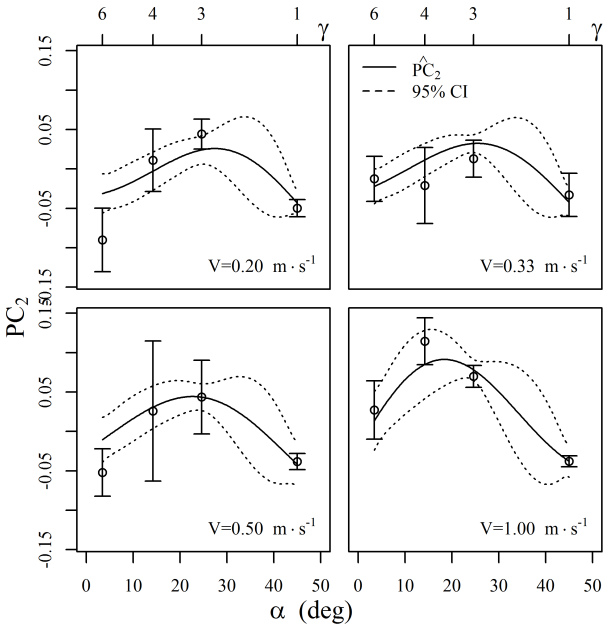


Fig. 14: Mean PC_2 evolution versus α and the corresponding GP model prediction and 95% confidence bounds. Error bars correspond to the bootstrapped variation for PC_2.

distances. With increasing strain (decreasing rake angle), there is less autocorrelation at large length scales but correlations in the intermediate values $(0.5 - 2\,\mu m)$ persist and hence PC_2 increases. However, this trend reverses at the lowest rake angles (highest strains) when the autocorrelation becomes extremely localized ($< 500\,nm$) and thus less similarity is observed in the $0.5 - 2\,\mu m$ range. These interactions are complex because each basis captures several *coupled* physical features (e.g. PC_2 captures negative long range and positive medium range autocorrelation). Furthermore, the bases must interact and balance their respective contributions in order to describe the changing physics at different machining process settings.

Figure 17 illustrates that the Vickers hardness and indentation yield produce similar trends with respect to the rake angle. For reference the mean virgin material hardness is $HV = 87.5 \pm 5.0$ (95% confidence interval). At large rake angles (low strains) the generated chips have higher hardness than the virgin material but produce lower range properties relative to measurements at small rake angles (larger strains). This observation is in accordance with deformation induced strain hardening. For cutting speeds $V = 0.20,\ 0.33,\ 0.50\,m \cdot s^{-1}$, the hardness appears to saturate with decreasing rake angle, which indicates that additional straining does not drive an increase in hardness. At the lowest cutting speed, however, indentation yield produces a fairly linear

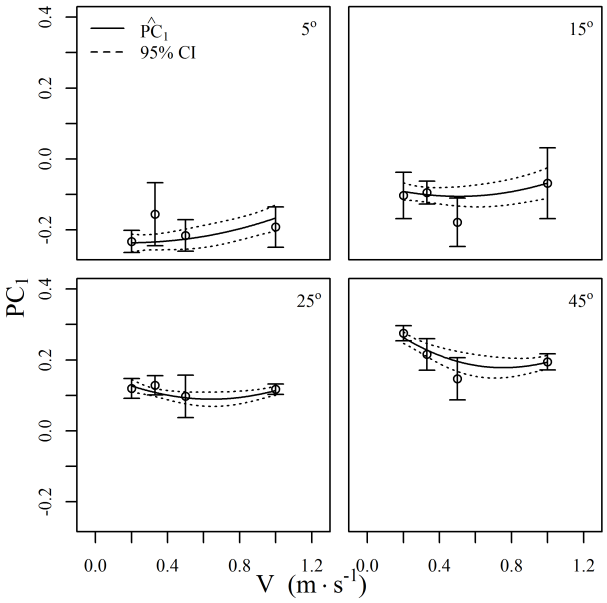


Fig. 15: Mean PC_1 evolution versus V and the corresponding GP model prediction and 95% confidence bounds. Error bars correspond to the bootstrapped variation for PC_1.

trend which decreased with increasing speed. Therefore, hardness and yield do not always share a one-to-one correspondence but nevertheless the inclusion of hardness is informative. At the highest speed and lowest rake angle (highest strain) there is a significant decrease in both hardness and strength. This is likely driven by recovery processes, which occur due to the higher cutting temperatures experienced under these conditions.

In this study we only consider structural morphology and therefore neglect crystallographic effects. This is one potential source of the scatter observed in Figure 17. The local crystallographic orientation of the indented site will likely influence the indentation response. However, for simplicity we adopt a strategy where this was neglected and instead homogenized over many observations. When crystallographic information is desirable, the stand alone GSH representation (which quantifies the ODF) may be augmented as additional features to $\bar{p}_{\boldsymbol{t}}$. Another possibility is to use the strategy established in [43,44] and use the paired two point statistics between each of the GSH coefficients. Recall that the GSH representation is a sum over multiple indices (μ, n, l) and in this work we truncate to 10 terms. Each of these 10 terms can be used as a measure of microstructural state. Therefore, these state descriptors may be used to compute two point spatial correlations [43]. Including constraints and symmetry considerations, it may be shown that there are $2 \cdot 10 - 1 = 19$

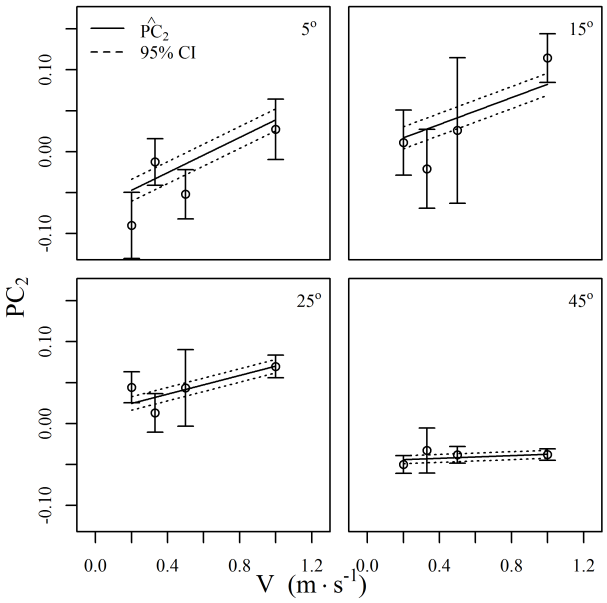


Fig. 16: Mean PC_2 evolution versus V and the corresponding GP model prediction and 95% confidence bounds. Error bars correspond to the bootstrapped variation for PC_2.

unique correlation pairs [43]. The derived expression in Eqn. 4 happens to be the mean over all auto-correlation pairs considered in [43]. The derivation in this work is fairly compact and proves that this mean quantity has a physical interpretation and is a descriptor of morphological spatial crystallographic "spread", which includes misorientation.

Bootstrapping methodology appears to be an effective method for quantifying the dispersion of microstructure in reduced order PC space, as shown in Figure 10. In our regression model bootstrapping is useful as it eliminates the need to estimate the measurement error variances when training the MOGPR model – instead they can be estimated directly from bootstrapping. Note however that bootstrapping of correlated data requires that the original sample be sufficiently large such that it "contains" the relevant correlation length scales. In our setting, the correlation length scales, particularly at large rake angles (low strain), are larger than the image field of view. Nevertheless, the bootstrapped variance estimates will reflect this artifact; inadequately sized images will yield more variance. Additionally, the disparity in autocorrelation at the lower spatial length scales is sufficiently significant that trends are still clear despite "missing" information at very large length scales.

The MOGPR model is effective at quantifying PSP relationships and provides estimates for coupled structure-property uncertainties. A natural con-

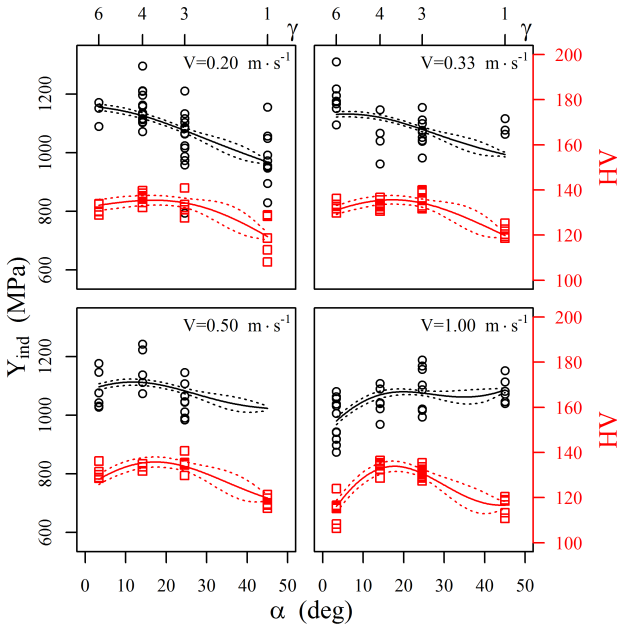


Fig. 17: Mean Y_{ind} and HV evolution versus α and the corresponding GP model prediction and 95% confidence bounds.

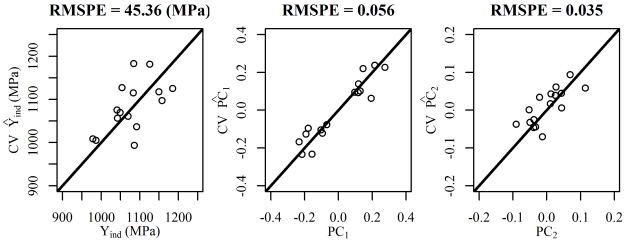
cern however is that perhaps the obtained hyperparameters, $\hat{\boldsymbol{\Phi}}$ in Eqn. A.8, neglect structure-property relationships. In Eqn. A.6 the structure-property linkage is captured through the cross-correlation matrix $\boldsymbol{S}$, which must be inferred from the observed data. This matrix quantifies the covariance (or correlation) between all the outputs considered $(PC_1, PC2, Y_{ind})$. The case where structure-property linkages are neglected the covariance matrix would take a block form,

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{C}_{11} + \boldsymbol{\Sigma}_{11} & \boldsymbol{C}_{21} + \boldsymbol{\Sigma}_{21} & \cdots & \mathbf{0} \\ \boldsymbol{C}_{12} + \boldsymbol{\Sigma}_{12} & \boldsymbol{C}_{22} + \boldsymbol{\Sigma}_{22} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \boldsymbol{C}_M + \boldsymbol{\Sigma}_M \end{bmatrix}, \tag{5}$$

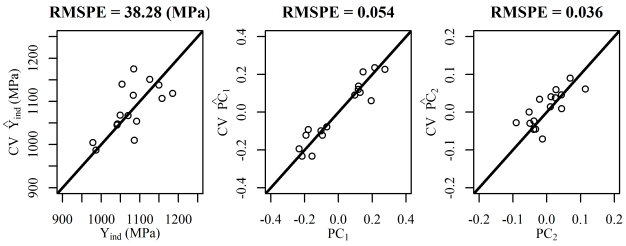
which suggests no correlation between the PC's and Y_{ind}. This degenerate case corresponds to two *independent* Gaussian process models; one for the process-structure and another for process-properties. Yet another degenerate case corresponds to a diagonal covariance structure where no correlation exists between any of the considered variables and thus the result is M independent GP models. However, consider that this model is data-driven and therefore it is *possible* that perhaps a process-property relationship does exist. In fact, there is some recent evidence in the literature that suggests that these map-

pings are plausible in some settings [3]. The inclusion of structural information is physically motivated and is expected to yield better performance as the data is much *richer* if structure information is included. The merit of the MOGPR model is that all possibilities may be considered at once; if a direct process-property linkage exists then the model will identify it. Note that it may seem inappropriate to assume structure-structure cross-correlations between the PC weights as PCA theory generates PC weights which are independent e.g. $\mathrm{Cov}\left(PC_1, PC_2\right) = 0$. However, this is only true in the unsupervised setting; PC weights are independent when nothing is known about the process settings. The PC basis and weights are computed from the unlabeled $p_{\boldsymbol{t}}$ ensemble of observations. In the MOGPR model correlation between PC_1 and PC_2 is possible because the correlation is *conditional* on also knowing the process settings. Two uncorrelated random variables may become correlated when conditioned on a third random variable related to the first two. Clearly, in the second case the two otherwise independent experiments become correlated due to the extra information.

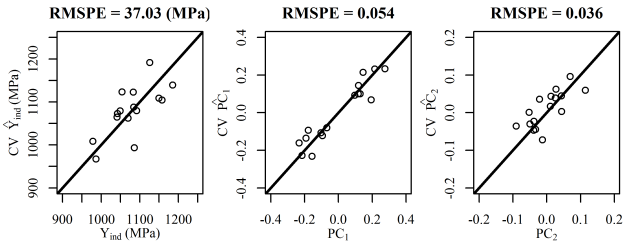
Cross validation results using a leave-one-unique-process-setting-out strategy are displayed in Figure 18. The cross validation results may be exactly computed from the fully trained model by employing a short-cut formula; see Appendix C. Four different results are shown to illustrate a few key points: (a) cross validation using Y_{ind} property data not including output cross-correlations (the case of M independent GP models), (b) cross validation using Y_{ind} property data with structure-property cross correlations, (c) cross validation using Y_{ind} and HV as coupled properties with no structure-property cross-correlations, and finally (d) cross validation considering all available property data (Y_{ind} and HV) and including output cross-correlations. Notice that strategy (d), which considers all property data and all correlations, yields the best cross validation error (25% improvement in Y_{ind} prediction relative to model (a)). Therefore inclusion of the hardness data did improve the overall model performance. Furthermore, each increase in model complexity provides slight improvements over the previous model. In general, however, this may not always be the case. GPR models are also prone to over-fitting when there is an imbalance between model complexity and data. For this reason some researchers prefer to use cross validation strategies for model training [45].



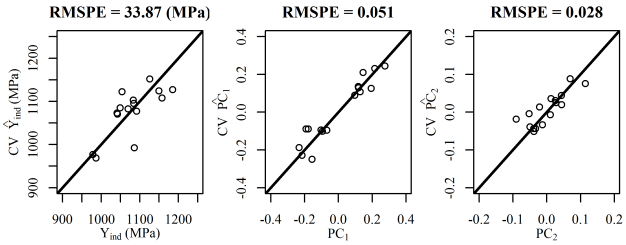
(a) Model only considering Y_{ind} as a property measure with no cross correlations (e.g. direct process-property and process-structure models)



(b) Model only considering Y_{ind} as a property measure with cross correlations



(c) Multi-fidelity model including Y_{ind} and HV with no cross correlations (e.g. direct process-property and process-structure models)



(d) Multi-fidelity model including Y_{ind} and HV as well as structure-property cross correlations

Fig. 18: Cross validation results removing one unique process setting at a time.

6 Conclusions

In this work we studied a severe plastic deformation machining process which drives microstructure evolution via continuous dynamic recrystallization. Various stages of microstructure evolution were captured by considering a wide range of rake angles, which induce a wide range of shear strains. Rate and temperature effects were considered by varying the cutting speed. Large strain conditions produced sub-micron crystal structures whereas low strain experiments yielded highly deformed structures, which still resembled the coarse parent material. At the largest strains a dependence on the cutting speed was observed with higher cutting speeds producing structures with lower crystallographic misorientations. Generalized spherical harmonics were used to efficiently quantify the local orientation state and a novel autocorrelation spatial statistic was derived that captures orientation "spread" or misorientation. The novel descriptor is physically intuitive and targets morphological information present in the orientation imaging data. A data driven multiple output Gaussian process regression model was established for quantifying process-structure-property linkages. The model is flexible, enables inclusion of various kinds of structure and property data, does not necessitate fully paired input data, captures the full process-structure-property pipeline, and produces coupled uncertainty estimates associated with future predictions.

7 Acknowledgements

The authors are grateful to the Woodruff School machine shop for their assistance in manufacturing of the custom cutting tool used in this work. Financial support of the work by the Morris M. Bryan, Jr. Professorship is acknowledged.

8 Conflicts of Interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Abolghasem, S., Basu, S., Shekhar, S., Cai, J., Shankar, M.: Mapping subgrain sizes resulting from severe simple shear deformation. Acta Materialia **60**(1), 376–386 (2012)
2. Adams, B.L., Gao, X.C., Kalidindi, S.R.: Finite approximations to the second-order properties closure in single phase polycrystals. Acta Materialia **53**(13), 3563–3577 (2005)
3. Agrawal, A., Deshpande, P.D., Cecen, A., Basavarsu, G.P., Choudhary, A.N., Kalidindi, S.R.: Exploration of data science techniques to predict fatigue strength of steel from composition and processing parameters. Integrating Materials and Manufacturing Innovation **3**(1), 8 (2014)
4. Basu, S., Shankar, M.R.: Crystallographic textures resulting from severe shear deformation in machining. Metallurgical and Materials Transactions A **46**(2), 801–812 (2015)

5. Basu, S., Wang, Z., Liu, R., Saldana, C.: Enhanced subsurface grain refinement during transient shear-based surface generation. Acta Materialia **116**, 114–123 (2016)
6. Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg (2006)
7. Boyle, P., Frean, M.: Dependent gaussian processes. In: Advances in neural information processing systems, pp. 217–224 (2005)
8. Brown, T.L., Saldana, C., Murthy, T.G., Mann, J.B., Guo, Y., Allard, L.F., King, A.H., Compton, W.D., Trumble, K.P., Chandrasekar, S.: A study of the interactive effects of strain, strain rate and temperature in severe plastic deformation of copper. Acta Materialia **57**(18), 5491–5500 (2009)
9. Bunge, H.J.: Texture analysis in materials science: mathematical methods. Elsevier (2013)
10. Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A.: Stan: A probabilistic programming language. Journal of statistical software **76**(1) (2017)
11. Cecen, A., Fast, T., Kalidindi, S.R.: Versatile algorithms for the computation of 2-point spatial correlations in quantifying material structure. Integrating Materials and Manufacturing Innovation **5**(1), 1 (2016)
12. Cecen, A., Fast, T., Kumbur, E., Kalidindi, S.: A data-driven approach to establishing microstructure–property relationships in porous transport layers of polymer electrolyte fuel cells. Journal of Power Sources **245**, 144–153 (2014)
13. Cecen, A., Wargo, E., Hanna, A., Turner, D., Kalidindi, S., Kumbur, E.: 3-d microstructure analysis of fuel cell materials: spatial distributions of tortuosity, void size and diffusivity. Journal of The Electrochemical Society **159**(3), B299–B307 (2012)
14. Cecen, A., Yabansu, Y.C., Kalidindi, S.R.: A new framework for rotationally invariant two-point spatial correlations in microstructure datasets. Acta Materialia (2018)
15. Deshpande, P., Gautham, B., Cecen, A., Kalidindi, S., Agrawal, A., Choudhary, A.: Application of statistical and machine learning techniques for correlating properties to composition and manufacturing processes of steels. In: Proceedings of the 2nd World Congress on Integrated Computational Materials Engineering (ICME), pp. 155–160. Springer (2013)
16. Efron, B., Tibshirani, R.J.: An introduction to the bootstrap. CRC press (1994)
17. Fast, T., Niezgoda, S.R., Kalidindi, S.R.: A new framework for computationally efficient structure–structure evolution linkages to facilitate high-fidelity scale bridging in multi-scale materials models. Acta Materialia **59**(2), 699–707 (2011)
18. Fatemi-Varzaneh, S., Zarei-Hanzaki, A., Beladi, H.: Dynamic recrystallization in az31 magnesium alloy. Materials Science and Engineering: A **456**(1-2), 52–57 (2007)
19. Fernandez-Zelaia, P.: Machining psp. https://github.com/pfz3 (2019)
20. Fernandez-Zelaia, P., Joseph, V.R., Kalidindi, S.R., Melkote, S.N.: Estimating mechanical properties from spherical indentation using bayesian approaches. Materials & Design **147**, 92–105 (2018)
21. Fernandez-Zelaia, P., Melkote, S.N.: Statistical calibration and uncertainty quantification of complex machining computer models. International Journal of Machine Tools and Manufacture (2018)
22. Fuentes, M.: A high frequency kriging approach for non-stationary environmental processes. Environmetrics: The official journal of the International Environmetrics Society **12**(5), 469–483 (2001)
23. Guo, Y., Saldana, C., Compton, W.D., Chandrasekar, S.: Controlling deformation and microstructure on machined surfaces. Acta materialia **59**(11), 4538–4547 (2011)
24. Haaland, B., Qian, P.Z.: An approach to constructing nested space-filling designs for multi-fidelity computer experiments. Statistica Sinica **20**(3), 1063 (2010)
25. Hoff, P.D.: A first course in Bayesian statistical methods. Springer Science & Business Media (2009)
26. Ion, S., Humphreys, F., White, S.: Dynamic recrystallisation and the development of microstructure during the high temperature deformation of magnesium. Acta Metallurgica **30**(10), 1909–1919 (1982)
27. Iskakov, A., Yabansu, Y.C., Rajagopalan, S., Kapustina, A., Kalidindi, S.R.: Application of spherical indentation and the materials knowledge system framework to establishing microstructure-yield strength linkages from carbon steel scoops excised from high-temperature exposed components. Acta Materialia **144**, 758–767 (2018)

28. Kalidindi, S.R.: Hierarchical materials informatics: novel analytics for materials data. Elsevier (2015)
29. Kalidindi, S.R., Brough, D.B., Li, S., Cecen, A., Blekh, A.L., Congo, F.Y.P., Campbell, C.: Role of materials data science and informatics in accelerated materials innovation. Mrs Bulletin **41**(8), 596–602 (2016)
30. Kalidindi, S.R., Medford, A.J., McDowell, D.L.: Vision for data and informatics in the future materials innovation ecosystem. JOM **68**(8), 2126–2137 (2016)
31. Kennedy, M.C., O'Hagan, A.: Bayesian calibration of computer models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **63**(3), 425–464 (2001)
32. Khosravani, A., Cecen, A., Kalidindi, S.R.: Development of high throughput assays for establishing process-structure-property linkages in multiphase polycrystalline metals: Application to dual-phase steels. Acta Materialia **123**, 55–69 (2017)
33. Me-Bar, Y., Shechtman, D.: On the adiabatic shear of ti 6al 4v ballistic targets. Materials Science and Engineering **58**(2), 181–188 (1983)
34. Minnaar, K., Zhou, M.: An analysis of the dynamic shear failure resistance of structural metals. Journal of the Mechanics and Physics of Solids **46**(10), 2155–2170 (1998)
35. M'Saoubi, R., Larsson, T., Outeiro, J., Guo, Y., Suslov, S., Saldana, C., Chandrasekar, S.: Surface integrity analysis of machined inconel 718 over multiple length scales. CIRP Annals-Manufacturing Technology **61**(1), 99–102 (2012)
36. Murr, L., Ramirez, A., Gaytan, S., Lopez, M., Martinez, E., Hernandez, D., Martinez, E.: Microstructure evolution associated with adiabatic shear bands and shear band failure in ballistic plug formation in ti–6al–4v targets. Materials Science and Engineering: A **516**(1-2), 205–216 (2009)
37. Ni, H., Elmadagli, M., Alpas, A.: Mechanical properties and microstructures of 1100 aluminum subjected to dry machining. Materials Science and Engineering: A **385**(1-2), 267–278 (2004)
38. Niezgoda, S.R.: Stochastic representation of microstructure via higher-order statistics: theory and application (2010)
39. Niezgoda, S.R., Yabansu, Y.C., Kalidindi, S.R.: Understanding and visualizing microstructure and microstructure variance as a stochastic process. Acta Materialia **59**(16), 6387–6400 (2011)
40. Pan, S.J., Yang, Q., et al.: A survey on transfer learning. IEEE Transactions on knowledge and data engineering **22**(10), 1345–1359 (2010)
41. Pathak, S., Shaffer, J., Kalidindi, S.R.: Determination of an effective zero-point and extraction of indentation stress–strain curves without the continuous stiffness measurement signal. Scripta Materialia **60**(6), 439–442 (2009)
42. Pathak, S., Stojakovic, D., Doherty, R., Kalidindi, S.R.: Importance of surface preparation on the nano-indentation stress-strain curves measured in metals. Journal of Materials Research **24**(3), 1142–1155 (2009)
43. Paulson, N.H., Priddy, M.W., McDowell, D.L., Kalidindi, S.R.: Reduced-order structure-property linkages for polycrystalline microstructures based on 2-point statistics. Acta Materialia **129**, 428–438 (2017)
44. Paulson, N.H., Priddy, M.W., McDowell, D.L., Kalidindi, S.R.: Data-driven reduced-order models for rank-ordering the high cycle fatigue performance of polycrystalline microstructures. Materials & Design (2018)
45. Pilania, G., Mannodi-Kanakkithodi, A., Uberuaga, B., Ramprasad, R., Gubernatis, J., Lookman, T.: Machine learning bandgaps of double perovskites. Scientific reports **6**, 19375 (2016)
46. Rasmussen, C.E.: Gaussian processes in machine learning. In: Advanced lectures on machine learning, pp. 63–71. Springer (2004)
47. Sagapuram, D., Yeung, H., Guo, Y., Mahato, A., M'Saoubi, R., Compton, W.D., Trumble, K.P., Chandrasekar, S.: On control of flow instabilities in cutting of metals. CIRP Annals **64**(1), 49–52 (2015)
48. Santner, T.J., Williams, B.J., Notz, W.I.: The design and analysis of computer experiments. Springer Science & Business Media (2013)
49. Shaw, M.C., Cookson, J.: Metal cutting principles. Clarendon press Oxford (1984)
50. Shaw, M.C., Cookson, J.: Metal cutting principles, vol. 2. Oxford university press New York (2005)

51. Shekhar, S., Abolghasem, S., Basu, S., Cai, J., Shankar, M.: Effect of severe plastic deformation in machining elucidated via rate-strain-microstructure mappings. Journal of Manufacturing Science and Engineering **134**(3), 031008 (2012)
52. Stan Development Team: RStan: the R interface to Stan (2018). URL http://mc-stan.org/. R package version 2.17.3
53. Sundararaghavan, V., Zabaras, N.: A dynamic material library for the representation of single-phase polyhedral microstructures. Acta Materialia **52**(14), 4111–4119 (2004)
54. Sundararaghavan, V., Zabaras, N.: Classification and reconstruction of three-dimensional microstructures using support vector machines. Computational Materials Science **32**(2), 223–239 (2005)
55. Swaminathan, S., Shankar, M.R., Lee, S., Hwang, J., King, A.H., Kezar, R.F., Rao, B.C., Brown, T.L., Chandrasekar, S., Compton, W.D., et al.: Large strain deformation and ultra-fine grained materials by machining. Materials Science and Engineering: A **410**, 358–363 (2005)
56. Torquato, S.: Random heterogeneous materials: microstructure and macroscopic properties, vol. 16. Springer Science & Business Media (2013)
57. Tóth, L., Beausir, B., Gu, C., Estrin, Y., Scheerbaum, N., Davies, C.: Effect of grain refinement by severe plastic deformation on the next-neighbor misorientation distribution. Acta Materialia **58**(20), 6706–6716 (2010)
58. Tuo, R., Wu, C.J., Yu, D.: Surrogate modeling of computer experiments with different mesh densities. Technometrics **56**(3), 372–380 (2014)
59. Valiev, R.Z., Islamgaliev, R.K., Alexandrov, I.V.: Bulk nanostructured materials from severe plastic deformation. Progress in materials science **45**(2), 103–189 (2000)
60. Wang, Z., Basu, S., Murthy, T.G., Saldana, C.: Gradient microstructure and texture in wedge-based severe plastic burnishing of copper. Journal of Materials Research **33**(8), 1046–1056 (2018)
61. Wang, Z., Basu, S., Saldana, C.: Low-temperature machining in a fully submerged cryogenic environment. Machining Science and Technology **21**(1), 19–36 (2017)
62. Wargo, E., Hanna, A., Cecen, A., Kalidindi, S., Kumbur, E.: Selection of representative volume elements for pore-scale analysis of transport in fuel cell materials. Journal of power sources **197**, 168–179 (2012)
63. Yabansu, Y.C., Patel, D.K., Kalidindi, S.R.: Calibrated localization relationships for elastic response of polycrystalline aggregates. Acta Materialia **81**, 151–160 (2014)
64. Yang, Z., Li, X., Brinson, L.C., Choudhary, A.N., Chen, W., Agrawal, A.: Microstructural materials design via deep adversarial learning methodology. arXiv preprint arXiv:1805.02791 (2018)
65. Zhilyaev, A.P., Langdon, T.G.: Using high-pressure torsion for metal processing: Fundamentals and applications. Progress in Materials Science **53**(6), 893–979 (2008)

A MOGPR implementation

Consider a process who's input are $\boldsymbol{x}$ and that has multiple outputs $[Y_1, Y_2, \ldots, Y_K]$ which are observed with some measurement error $\boldsymbol{\epsilon}$. This process can be modeled using a multivariate GP model,

$$\begin{pmatrix} Y_1(\boldsymbol{x}) \\ Y_2(\boldsymbol{x}) \\ \vdots \\ Y_M(\boldsymbol{x}) \end{pmatrix} \sim \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{x}), \boldsymbol{S} + \mathrm{Cov}\boldsymbol{\epsilon}) \qquad (A.6)$$
$$S_{ij} = \mathrm{Cov}(Y_i, Y_j),$$

where the mean behavior of the outputs varies according to a mean *function* $\boldsymbol{\mu}(\boldsymbol{x})$, cross-correlation of outputs are captured through $\boldsymbol{S}$, and the observation errors are in general allowed to be correlated and perhaps have different scales for each outputs e.g. Cov$\boldsymbol{\epsilon}$ is purposefully generic. The mean function may be described using a parametric regression

strategy e.g. $\boldsymbol{\mu}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x})\boldsymbol{\beta}$ where $\boldsymbol{\beta}$ are regression coefficients and $\boldsymbol{f}(\boldsymbol{x})$ a vector of regressors. In this work we utilized linear and cross-linear terms (4 terms including constant) for each output Y_i. Note that we implicitly assume that the cross-correlations are spatially invariant and therefore stationary. Strategies exist for introducing non-stationarity [22,7] and we have successfully utilized these for developing FE surrogates however in this work we will utilize the simpler stationary cross-correlation structure [21]. The codes provided online however include additional non-stationary complexity [19].

Now consider that observations of each output Y_i are made at $\boldsymbol{x}_{ij}$ where $j = 1, \ldots, N_i$. This notation is flexible enough to allow each Y_i output to have N_i unique observations $\boldsymbol{x}_{ij}$ with a total number of $N = \sum_{i=1}^{K} N_i$. Again this is valuable if the Y_i have different costs associated with obtaining them (optical vs TEM). The accumulated dataset therefore can be expressed as another multivariate normal,

$$\begin{pmatrix} Y_1\left(\boldsymbol{x}_{11}\right) \\ \vdots \\ Y_1\left(\boldsymbol{x}_{1N_1}\right) \\ Y_2\left(\boldsymbol{x}_{21}\right) \\ \vdots \\ Y_2\left(\boldsymbol{x}_{2N_2}\right) \\ \vdots \\ Y_M\left(\boldsymbol{x}_{M_1}\right) \\ \vdots \\ Y_M\left(\boldsymbol{x}_{MN_M}\right) \end{pmatrix} \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{C} + \boldsymbol{\Sigma}\right) \tag{A.7}$$

$$\mathrm{Cov}\left[Y_i\left(\boldsymbol{x}_{ik}\right), Y_j\left(\boldsymbol{x}_{jl}\right)\right] = S_{ij} R\left(\boldsymbol{x}_{ik} - \boldsymbol{x}_{jl}\right) + \sigma_{ij}\delta_{ij}\delta_{kl} + \sigma_{ij}\delta_{kl}$$

$$R(\boldsymbol{h}) = \exp\left(-\sum_{i=1}^{d} \phi_i h_i^2\right),$$

where R is the Gaussian correlation function, $\boldsymbol{\Sigma}$ is the total error covariance matrix, and $\boldsymbol{\phi}_i$ are the correlation length scales for each of the d dimensions of $\boldsymbol{x}$. Note that $\boldsymbol{\Sigma}$ is comprised of σ_{ij} and contains some flexibility for different kinds of experiments. Observations have some measurement variance σ_{ij} when outputs are identical ($i = j$) and are observed at the same $\boldsymbol{x}$ ($k = l$) however if there are *paired* observations (Y_i, Y_j) at each observation ($k = l$) then there may also be correlations in the errors. One example where this may be relevant is when considering the PC-weights as microstructure descriptors which will generate pairs (or tuples in higher dimensions) of data for each micrograph. If observations are not measured in pairs then they should be independent and share no correlation. Note that if all experiments are performed at the same $\boldsymbol{x}$ then the above covariance structure has a Kronecker structure which can be exploited for computational efficiency [20]. The covariance structure imposes that there is a distance-based criteria for quantifying correlations (R), that there are cross-correlations across outputs ($\boldsymbol{\Sigma}$), and that there is a random component associated with measurement uncertainty ($\boldsymbol{\Sigma}$).

The *prior* placed on the data is that observations can be explained by interpreting them as coming from some multivariate normal generating process. As such the unknown hyperparameters $\boldsymbol{\Phi} = [\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{S}, \boldsymbol{\Sigma}]$ can be obtained from a maximum likelihood estimate (MLE) or a maximum a posteriori (MAP) estimate if priors are placed on some hyperparameters. The negative log-likelihood (or in a Bayesian setting the posterior) may be minimized to obtain estimates,

$$\hat{\boldsymbol{\Phi}} = \arg\min_{\boldsymbol{\Phi}} \left(\log|\boldsymbol{C} + \boldsymbol{\Sigma}| + (\boldsymbol{Y} - \boldsymbol{\mu})^T (\boldsymbol{C} + \boldsymbol{\Sigma})^{-1} (\boldsymbol{Y} - \boldsymbol{\mu})\right) \tag{A.8}$$

The statistical model was built in *Stan* [10], a statistical programming language, and evaluated using *RStan* [52] the *R*-language interface for *Stan*. The *Stan* optimizer was used to identify the MAP estimates of the hyperparameters $\hat{\boldsymbol{\Phi}}$. Note that in the inference there

are two matrix quantities that need to be estimated. To improve efficiency of the estimation and maintain that these matrices remain positive definite inverse-Wishart priors were placed on the matrix quantities and the inference was reparameterized. A review of this strategy can be found in [25].

Note that in Section 3.2 a methodology for bootstrapping confidence bounds on the mean PC-weights is presented. Therefore for each unique process setting we can establish the mean and variance measures of the mean PC-weights. The mean estimates should be used in constructing the data-vector, $\boldsymbol{Y}$, and the bootstrapped variance estimates can be prescribed to build the structure-portion of the error covariance matrix $\boldsymbol{\Sigma}$. This methodology also allows for inclusion of heteroskedastic variance estimates, which according to Figure 10, are appropriate. Furthermore, this rearrangement alleviates some computational burden associated with estimating some hyperparameters and enables use of bootstrapped quantities.

Predictions using a tuned MOGPR model can be easily obtained by again considering that the desired quantities, $(Y_1(\boldsymbol{x}), \ldots, Y_M(\boldsymbol{x}))^T$, are jointly distributed with the observed data

$$\begin{pmatrix} \boldsymbol{Y}_1 \\ \vdots \\ \boldsymbol{Y}_M \\ Y_1(\boldsymbol{x}) \\ \vdots \\ Y_M(\boldsymbol{x}) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu}(\boldsymbol{x}) \end{pmatrix} \begin{bmatrix} \boldsymbol{C}+\boldsymbol{\Sigma} & \boldsymbol{r} \\ \boldsymbol{r}^T & \boldsymbol{S} \end{bmatrix} \right) \tag{A.9}$$

$$\boldsymbol{r}_{ij} = \Sigma_{ij}\left[R\left(\boldsymbol{x}-\boldsymbol{x}_{i1}\right), \ldots, R\left(\boldsymbol{x}-\boldsymbol{x}_{iN_j}\right)\right]^T,$$

where $\boldsymbol{r}$ captures the spatial and cross-correlations of each output $Y_i(\boldsymbol{x})$ with all previously observed data $j = 1, \ldots, M$. The expectation of $(Y_1(\boldsymbol{x}), \ldots, Y_M(\boldsymbol{x}))^T$ conditional on all previous observations $(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_M)^T$ is

$$\begin{pmatrix} Y_1(\boldsymbol{x}) \\ \vdots \\ Y_M(\boldsymbol{x}) \end{pmatrix} = \boldsymbol{\mu}(\boldsymbol{x}) - \boldsymbol{r}^T\left(\boldsymbol{C}+\boldsymbol{\Sigma}\right)^{-1}\left(\boldsymbol{\mu}-\boldsymbol{Y}\right), \tag{A.10}$$

and the covariance associated with this prediction is,

$$\hat{\boldsymbol{S}} = \boldsymbol{S} - \boldsymbol{r}^T\left(\boldsymbol{C}+\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{r}. \tag{A.11}$$

From these two expressions it is clear that the MOGPR provides estimates for the PSP linkages as well as uncertainty predictions through the prediction covariance $\hat{\boldsymbol{S}}$.

B Multifidelity implementation

Statistically we can build a simple model that allows for sharing of information between the physically informative quantities (Y_{ind}) and the cheaper less informative quantities (HV)

$$\begin{aligned} Y_{ind} &= Z + \tau \\ HV &= \rho Z + W + \gamma, \end{aligned} \tag{B.12}$$

where Z is the underlying mean function we seek described as a Gaussian process (GP), τ is error associated with Y_{ind}, ρ is a scaling quantity, W is an independent zero-mean GP which allows HV to vary from ρZ systematically (e.g. bias function) and γ is the measurement error in HV. This form is identical to the form introduced in the seminal Kennedy and O'Hagan paper [31]. Note that Z is part of the multivariate GP previously introduced but we simply denote it here as Z for simplicity. The model states that HV scales with Z (and

hence the mean of Y_{ind}) except when the simple scaling fails in which case W "captures" or "soaks up" this deviation.

The covariance of HV with the other MOGPR quantities can be easily derived. First assume that $Y_{ind} = Y_{M-1}$ in the model e.g. the indentation yield is ordered as the second to last output, and HV is the last $Y_M = HV$. Therefore,

$$\begin{aligned}\mathrm{Cov}[Y_{ind}\left(\boldsymbol{x}_{M-1,k}\right), \;\; HV\left(\boldsymbol{x}_{Ml}\right)] &= \mathrm{Cov}[Z\left(\boldsymbol{x}_{M-1,k}\right) + \tau, \\ &\quad \rho Z\left(\boldsymbol{x}_{Ml}\right) + W\left(\boldsymbol{x}_{Ml}\right) + \gamma] \\ &= \rho S_{M-1,M-1} R\left(\boldsymbol{x}_{M-1,k} - \boldsymbol{x}_{Ml}\right),\end{aligned} \tag{B.13}$$

where σ_b^2 and R_b are the bias variance and correlation function. The bias correlation function contains additional hyperparameters $\boldsymbol{\phi}_b$. Note there is no error term since there are no off-diagonal terms in the error covariance structure. This is because indentation-hardness experiments are not "paired" experimentally; observations are made independently of one another. All other correlations can be easily obtained simply by following the above "plug-in" strategy. The bias function "kicks in" only for $HV - HV$ covariances,

$$\begin{aligned}\mathrm{Cov}[HV\left(\boldsymbol{x}_{Mk}\right), \;\; HV\left(\boldsymbol{x}_{Ml}\right)] &= \rho^2 S_{M-1,M-1} R\left(\boldsymbol{x}_{Mk} - \boldsymbol{x}_{Ml}\right) \\ &\quad + \sigma_b^2 R_b\left(\boldsymbol{x}_{Mk} - \boldsymbol{x}_{Ml}\right) + \\ &\quad \gamma\delta_{kl},\end{aligned} \tag{B.14}$$

C Cross-validation shortcut formulas

The cross validation error associated with removing a subset of data of size N_i, represented by multi-index i, can be expressed as

$$\boldsymbol{cv}_i = \boldsymbol{y}_i - \hat{\boldsymbol{f}}_{(i)}(\boldsymbol{x}_i), \tag{C.15}$$

where $\boldsymbol{y}_i$ are the responses corresponding to i and $\hat{\boldsymbol{f}}_{(i)}(\boldsymbol{x}_i)$ is the corresponding estimate for a model which is trained by withholding data belonging to i. Note that $\boldsymbol{cv}_i$ is a vector with multiple observations of a potentially multivariate output (e.g. PC_1, PC_2, Y_{ind}). The model estimate is given by

$$\hat{\boldsymbol{f}}_{(i)}(\boldsymbol{x}_i) = \boldsymbol{\mu}(\boldsymbol{x}_i) + \widetilde{\boldsymbol{r}}_{(i)}^T \boldsymbol{C}_{(i)}^{-1}\left(\boldsymbol{y}_{(i)} - \boldsymbol{\mu}_{(i)}\right) \tag{C.16}$$

where quantities containing subscript (i) represent quantities computed with data from i missing.

Now consider the complete covariance matrix where the ordering of the data is rearranged such that the block corresponding to i is shifted to the final rows/columns,

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{C}_{(i)} & \widetilde{\boldsymbol{r}}_{(i)} \\ \widetilde{\boldsymbol{r}}_{(i)}^T & \widetilde{\boldsymbol{\Sigma}}_i \end{bmatrix}, \tag{C.17}$$

To compute $\boldsymbol{C}^{-1}$ the Sherman-Morrison-Woodbury formula can be applied,

$$\boldsymbol{C}^{-1} = \begin{bmatrix} \boldsymbol{C}_{(i)}^{-1} + \boldsymbol{B}\widetilde{\boldsymbol{r}}_{(i)}^T \boldsymbol{C}_{(i)}^{-1} & -\boldsymbol{B} \\ -\boldsymbol{B}^T & \left(\widetilde{\boldsymbol{\Sigma}}_i - \widetilde{\boldsymbol{r}}_{(i)}^T \boldsymbol{C}_{(i)}^{-1} \widetilde{\boldsymbol{r}}_{(i)}\right)^{-1} \end{bmatrix}, \tag{C.18}$$

where $\boldsymbol{B} = \boldsymbol{C}_{(i)}^{-1} \widetilde{\boldsymbol{r}}_{(i)} \left(\widetilde{\boldsymbol{\Sigma}}_i - \widetilde{\boldsymbol{r}}_{(i)}^T \boldsymbol{C}_{(i)}^{-1} \widetilde{\boldsymbol{r}}_{(i)}\right)^{-1}$. Note that this manipulation enables the interpretation of quantity $\left(\widetilde{\boldsymbol{\Sigma}}_i - \widetilde{\boldsymbol{r}}_{(i)}^T \boldsymbol{C}_{(i)}^{-1} \widetilde{\boldsymbol{r}}_{(i)}\right)^{-1}$ as the i^{th} "block-diagonal" entry of $\boldsymbol{C}^{-1}$. Here we are referring

to the $N_i \times N_i$ entry corresponding to indices i which will be noted as $\boldsymbol{C}_{ii}^{-1}$. Similarly $-\left(\widetilde{\boldsymbol{\Sigma}}_i - \widetilde{\boldsymbol{r}}_{(i)}^T \boldsymbol{C}_{(i)}^{-1} \widetilde{\boldsymbol{r}}_{(i)}\right)^{-1} \widetilde{\boldsymbol{r}}_{(i)}^T \boldsymbol{C}_{(i)}^{-1}$ is the i^{th} "block-row" of $\boldsymbol{C}^{-1}$. Note this is really a $N_i \times (N_{tot} - N_i)$ matrix, where N_{tot} is the total number of data points, but we refer to it as a "block-row" because of its association with the i^{th} rows of the correlation matrix. This quantity will be noted $\boldsymbol{C}_{i,(i)}^{-1}$ e.g. the i^{th} block-row of $\boldsymbol{C}^{-1}$ not including the i^{th} block-diagonal portion. Therefore

$$\begin{aligned} \boldsymbol{C}_{i,(i)}^{-1} &= -\left(\widetilde{\boldsymbol{\Sigma}}_i - \widetilde{\boldsymbol{r}}_{(i)}^T \boldsymbol{C}_{(i)}^{-1} \widetilde{\boldsymbol{r}}_{(i)}\right)^{-1} \widetilde{\boldsymbol{r}}_{(i)}^T \boldsymbol{C}_{(i)}^{-1} \\ &= -\boldsymbol{C}_{ii}^{-1} \widetilde{\boldsymbol{r}}_{(i)}^T \boldsymbol{C}_{(i)}^{-1} \\ \widetilde{\boldsymbol{r}}_{(i)}^T \boldsymbol{C}_{(i)}^{-1} &= -\left(\boldsymbol{C}_{ii}^{-1}\right)^{-1} \boldsymbol{C}_{i,(i)}^{-1}. \end{aligned} \tag{C.19}$$

The advantage of these manipulations will become clear when returning to equations C.15 and C.16

$$\begin{aligned} \boldsymbol{cv}_i &= \boldsymbol{y}_i - \boldsymbol{\mu}(\boldsymbol{x}_i) - \widetilde{\boldsymbol{r}}_{(i)}^T \boldsymbol{C}_{(i)}^{-1} \left(\boldsymbol{y}_{(i)} - \boldsymbol{\mu}_{(i)}\right) \\ &= \boldsymbol{y}_i - \boldsymbol{\mu}(\boldsymbol{x}_i) + \left(\boldsymbol{C}_{ii}^{-1}\right)^{-1} \boldsymbol{C}_{i,(i)}^{-1} \left(\boldsymbol{y}_{(i)} - \boldsymbol{\mu}_{(i)}\right) \\ &= \boldsymbol{y}_i - \boldsymbol{\mu}(\boldsymbol{x}_i) + \\ &\quad \left(\boldsymbol{C}_{ii}^{-1}\right)^{-1} \left[\boldsymbol{C}_{i,(i)}^{-1} \; \boldsymbol{C}_{ii}^{-1}\right] \left(\begin{bmatrix} \boldsymbol{y}_{(i)} \\ \boldsymbol{y}_i \end{bmatrix} - \begin{bmatrix} \boldsymbol{\mu}_{(i)} \\ \boldsymbol{\mu}_i \end{bmatrix}\right) \\ &\quad - \left(\boldsymbol{C}_{ii}^{-1}\right)^{-1} \boldsymbol{C}_{ii}^{-1} \boldsymbol{y}_i + \left(\boldsymbol{C}_{ii}^{-1}\right)^{-1} \boldsymbol{C}_{ii}^{-1} \boldsymbol{\mu}_i \\ \boldsymbol{cv}_i &= \left(\boldsymbol{C}_{ii}^{-1}\right)^{-1} \boldsymbol{C}_i^{-1} \left(\boldsymbol{y} - \boldsymbol{\mu}\right) \\ \boldsymbol{cv} &= \mathit{blockdiag}\left(\boldsymbol{C}^{-1}\right)^{-1} \boldsymbol{C}^{-1} \left(\boldsymbol{y} - \boldsymbol{\mu}\right). \end{aligned} \tag{C.20}$$

In the above manipulations $\boldsymbol{\mu}(\boldsymbol{x}_i) = \boldsymbol{\mu}_i$, $\left[\boldsymbol{C}_{i,(i)}^{-1} \; \boldsymbol{C}_{ii}^{-1}\right] = \boldsymbol{C}_i^{-1}$ e.g. the i^{th} block-row, and $\left(\boldsymbol{C}_{ii}^{-1}\right)^{-1} \boldsymbol{C}_{ii}^{-1} = \boldsymbol{I}$ where $\boldsymbol{I}$ is the identity matrix. These manipulations enable the direct computation of the leave-i-out cross validation. Computationally this is a much more favorable estimate over the alternative which would require retraining many times for however many i there are. This methodology can easily be applied towards *k-folds* cross-validation where the previously introduced i would correspond indices in the matrix belonging to each of the k-folds.

An expression for the leave-one-out prediction variance can be obtained using the same matrix manipulations. For i^{th} hold out case the prediction covariance for $\hat{\boldsymbol{f}}_{(i)}(\boldsymbol{x}_i)$ can be obtained from,

$$\boldsymbol{S}_i = \left(\boldsymbol{C}_{ii}^{-1}\right)^{-1} \boldsymbol{C}_i^{-1} \boldsymbol{C}_i^T, \tag{C.21}$$

the diagonal of which contains the prediction variance estimates.