Title

Using machine-learning to understand complex microstructural effects on the mechanical behavior of Ti-6Al-4V alloys

Authors

McElfresh, Cameron
Roberts, Collin
He, Sicong
et al.

Publication Date

2022-06-01

DOI

10.1016/j.commatsci.2022.111267

Using machine-learning to understand complex microstructural effects on the mechanical behavior of Ti-6Al-4V alloys

Cameron McElfresh[a], Collin Roberts[a], Sicong He[a], Sergey Prikhodko[a], Jaime Marian[a]

[a]*Department of Materials Science and Engineering, University of California Los Angeles, Los Angeles, CA 90095, USA*

Abstract

Structural materials properties are highly dependent on their microstructure. Their microstructure is in turn affected by multiple fabrication and thermo-mechanical treatment parameters, all of which conform a highly-dimensional parametric space with often hidden correlations that are difficult to extract by experimentation alone. This is particularly true for alloys of the dual-phase Ti-6Al-4V family, with their greatly complex and rich microstructures, which combine several intrinsic lengthscales associated with multiple grain and subgrain structures, grains with different crystal lattices (α and β phases), and complex chemistry. Here we use a comprehensive set of machine learning techniques to develop predictive tools relating the yield strength and hardening rate of these alloys to a set of input parameters covering extensive ranges. The data generator is a finite-element crystal plasticity model for polycrystal deformation that takes into account slip anisotropy and employs standard dislocation evolution models for the α and β phases of Ti-based alloys. Our dataset includes over two thousand independent simulations and is used to train the machine learning models, which are then used to establish correlations between microstructural parameters and the alloys' mechanical response. Our results point to the most influential parameters affecting yield strength and hardening rate, information that can then be used to guide experimental synthesis and characterization efforts to save time and resources.

Keywords: Ti-6Al-4V alloys, Yield strength, Polycrystal plasticity, Dual-phase Ti alloys, Machine learning, Hardening rate

1. Introduction

Titanium alloys are widely used in aerospace, biomedical, transportation, and military applications due to their specific high strength and fracture toughness, corrosion resistance, and high-temperature properties. In particular, Ti-6Al-4V is one of the most popular titanium alloys due to the increased strength achieved with the stabilization of the body-centered cubic (BCC) β phase. The interplay between the hexagonal close-packed (HCP) α phase and the β phase is the primary factor dictating the strength of titanium alloys. Though the β phase is thermodynamically unfavorable below 890°C in pure titanium [1, 2], metastable dual-phase titanium alloys are possible through the addition of a mixture of beta-stabilizers (such as silicon and vanadium) and alpha-stabilizers (such as aluminum or oxygen). One of the primary advantages of α/β alloys is that they are heat treatable. Heat treating is vital to relieve residual stresses, remove machining history, or tune the microstructure for the ability of mass-manufacturing the material, making α/β alloys a continued area of interest for manufacturing and material science alike.

A number of previous studies have shown that thermo-mechanical processing of α/β Ti alloys can achieve a particular microstructure-performance combination for various applications [3]. Dual-phase titanium can exist in a number of microstructures including equiaxed, duplex, lamellar, intergranular, and lath [4, 5]. In addition, there is interest in better understanding graded or layered microstructures that may be ideal for lightweight or directionally-dependent high strength applications [6]. The advantageous properties of titanium alloys make them ideal candidates for many high-performance applications. However, the parametric space including all property-dependent microstructural features (grain size, phase fraction, orientation, texture, grain geometry, reinforcement particle size and distribution, solutes etc.) makes alloy design a multifactorial process of combinatorial proportions. As such, mapping the effect of one parameter –or of sets of different parameters– to a specific microstructural property or mechanical response through experimentation alone is thus impractical. For this reason, computational modeling and data analysis can become essential tools to establish direct correlations and narrow down the parametric space in search for improved alloys via micro- and macrostructural design.

Traditional computational modeling methods, such as crystal plasticity (CP), are ideal counterparts to experimental methods in order to study the various features associated with the deformation behavior and mechanical coupling between the β and α phases. Indeed, different variants of the CP method have been applied in recent times to modeling the deformation of dual-phase titanium alloys [7–11]. However, while useful to study specific aspects of microstructural evolution during alloy deformation, these tools alone cannot capture the complexities associated with the broad parametric space potentially influencing the material response. Capturing complex correlations between sets of variables requires using additional tools of statistical nature.

Advances in computing power and data availability have, among other things, propelled the widespread use of machine learning (ML) as an additional means of capturing meaning from data. The materials science modeling community has benefited from the use of machine learning techniques in studies utilizing density functional theory [12], dislocation dynamics [13], molecular dynamics [14], crystal plasticity [15], and others [16–18]. The relatively low cost to entry into the domain of machine learning makes it an ideal resource to complement data-heavy research processes. Moreover, machine learning techniques are ideal to apply when constructing a predictor for mechanical behavior because (i) there are often many features that affect a material's mechanical response, and (ii) the property-behavior relationship of the features tends to be non-linear. Many machine learning models excel at capturing non-linear behavior and have been successfully applied to build regressors that predict mechanical behavior for a wide range of materials including steels [19, 20], composites [21], and metallic glasses [22–24]. It should be emphasized that while there are constitutive equations that are used to model the mechanical behavior of alloys, the arbitrary extension of these expressions to include more (possibly non-linear) variables is not trivial [25, 26]. In this way, the use of ML as opposed to traditional constitutive expressions also decreases the rigor of expanding the model to include more feature variables as the data becomes available.

In this work, we employ several machine learning regression techniques in an exercise to develop predictive models for the strength and hardening rate of α/β dual-phase polycrystals. Our microstructure simulator is a crystal plasticity approach based on the work by Admal *et al.* [27] adapted to polycrystals with alternating BCC/HCP structures representative of dual-phase Ti alloys. The CP model is used to generate large data sets relating specific inputs to objective outputs, and ML regression techniques are then applied to assign importances and extract correlations.

The paper is organized as follows. In Section 2.1 we briefly outline the physical and computational background of the diffuse-crystal interface model. In Section 2.2 we describe the modifications made to the plasticity model in order

to simulate a dual-phase α/β titanium crystal and demonstrate acceptable correlation with literature values. In Section 3 we introduce the machine learning models. Section 4 presents the results of the simulations and machine learning regression exercise. In Section 5 we discuss the findings and potential next steps for improving the dual-phase model.

2. Numerical model

2.1. Diffuse Crystal Interface Plasticity Model

The following is a brief description of the previously-developed diffuse crystal interface plasticity model employed here. The original work can be found in ref. [27]. The basis of the diffuse-crystal interface model is to identify dislocations as the basic carrier of plastic deformation and build grain boundaries as continuum aggregates of these defects. In this fashion, grain boundaries are seen as incompatibilities of a plastic rotation field, which –much in the manner of standard elasto-plastic decompositions– must be closed by defining a special class of geometrically-necessary dislocations (GND) that habit the GB plane. Crystal deformation is modeled in the traditional sense, as a multiplicative combination of elastic and plastic deformations:

$$\boldsymbol{F}(\boldsymbol{X},t) = \boldsymbol{F}^{L}(\boldsymbol{X},t)\,\boldsymbol{F}^{P}(\boldsymbol{X},t) \tag{1}$$

where $\boldsymbol{F}^{L}$ and $\boldsymbol{F}^{P}$ are the lattice and plastic components of the total deformation gradient $\boldsymbol{F}$, respectively, at time t and position $\boldsymbol{X}$. The evolution of $\boldsymbol{F}^{P}$ is determined through the contribution of slip systems via slip rates using the flow rule:

$$\dot{\boldsymbol{F}}^{P} = \boldsymbol{L}^{P}\boldsymbol{F}^{P} \tag{2}$$

$\boldsymbol{L}^{P}$ is the plastic velocity gradient, defined as:

$$\boldsymbol{L}^{P}(\boldsymbol{X},t) := \sum_{\alpha=1^{N}} \gamma^{\alpha}\boldsymbol{s}^{\alpha}\otimes\boldsymbol{m}^{\alpha} \tag{3}$$

where $\boldsymbol{s}^{\alpha}$ and $\boldsymbol{n}^{\alpha}$ are unit vectors representing the glide and plane normal directions for slip system α. The value γ^{α} corresponds to the crystallographic slip rate on each slip system. The additional microscopic force and energy balance considerations are described in ref. [27]. Using standard crystal plasticity methods, a stress-free single crystal is constructed at $t = 0$ by requiring that:

$$\boldsymbol{F}^{L}(\boldsymbol{X},0) = \boldsymbol{F}^{P}(\boldsymbol{X},0) \equiv \boldsymbol{I} \tag{4}$$

such that $\boldsymbol{F} \equiv \boldsymbol{I}$. In contrast, the diffuse crystal interface model sets the initial state of the polycrystal to be:

$$\boldsymbol{F}^{L}(\boldsymbol{X},0) = \boldsymbol{R}^{0}(\boldsymbol{X}), \qquad \boldsymbol{F}^{P}(\boldsymbol{X},0) = \boldsymbol{R}^{0}(\boldsymbol{X})^{T} \tag{5}$$

where $\boldsymbol{R}^{0}$ represent the lattice rotation field in the polycrystal and maintains piecewise-constant values in each grain and smooth transitions across grain boundaries.

The rotational decomposition expressed in eq. (5) is the central framework to the diffuse-crystal interface plasticity model employed here. Using this decomposition, we can study polycrystals as a single boundary-value problem. Numerical discontinuities in $\boldsymbol{F}^{L}$ and $\boldsymbol{F}^{P}$ are avoided by implementing a smoothed step function in the space of the rotational fields. The remainder of the grain boundary and finite element numerical procedures remain the same as described in ref. [27]. However, the constitutive equations for plastic flow have been modified to accommodate the allotropic nature of α/β-Ti and the changes are described in Section 2.2.

2.2. Dislocation evolution model

The dual-phase nature of Ti-6Al-4V results in complex plastic deformation mechanisms that are not easily modeled. The microscopic force balance used here [27] is an extension of the approach developed by Barton *et al.* in [28] for BCC crystals, not applicable to the HCP α phase. For α-Ti, we adopt the model by Moore *et al.* [29], which is also based on a Kocks-Mecking dislocation density evolution law:

$$\dot{\rho} = \left(k_1\sqrt{\rho} - k_2\rho\right)\sum_{\alpha}^{N}|\dot{\gamma}^{\alpha}| \tag{6}$$

where ρ is the dislocation density, k_1 is the hardening parameter, k_2 is the recovery parameter, and $\sum_{\alpha}^{N} |\dot{\gamma}^{\alpha}|$ is the total shear rate. Limited hardening has been seen in near-α alloys in the elasto-plastic transient range [29], and thus we set k_2 to zero. The evolving dislocation density is used to calculate the slip system strength through:

$$g^{\alpha} = w^{\alpha} \left(g_0 + \tilde{\alpha} G b \sqrt{\rho}\right) \tag{7}$$

where g_0 is the lattice resistance, G is the shear modulus, b is the Burgers vector's modulus, and $\tilde{\alpha}$ is a material parameter that captures latent hardening. The anisotropy of the HCP crystal is embodied in the varying weights symbolized by the variable w^{α} which takes values of 1.0, 1.0, 1.1, and 3.0 for slip on the basal <a>, prismatic <a>, pyramidal <a>, and pyramidal <c+a>, respectively [30–32]. The inclusion of the slip system weight is one of the primary modifications to the plasticity model previously used in ref. [27] that enables extension to an allotropic HCP polycrystal. Lastly, the shear rates follow the standard strain-rate sensitivity dependence on stress, i.e.:

$$\dot{\gamma^{\alpha}} = \dot{\gamma}_0 \left| \frac{\tau^{\alpha}}{g^{\alpha}} \right|^{\frac{1}{m}} \text{sign}(\tau^{\alpha}) \tag{8}$$

where $\dot{\gamma}_0$ is a reference slip rate, τ^{α} is the resolved shear stress, g^{α} is the crystal strength, and m is the strain rate sensitivity exponent. τ^{α} is obtained as the Schmid projection on slip system α of the Cauchy stress, $\boldsymbol{\sigma}$:

$$\tau^{\alpha} = \boldsymbol{s}^{\alpha} \cdot \boldsymbol{\sigma} \cdot \boldsymbol{m}^{\alpha} \tag{9}$$

with:

$$\boldsymbol{\sigma} := \frac{C}{2} \left(\boldsymbol{F}^{T} \boldsymbol{F} - \boldsymbol{I} \right) \tag{10}$$

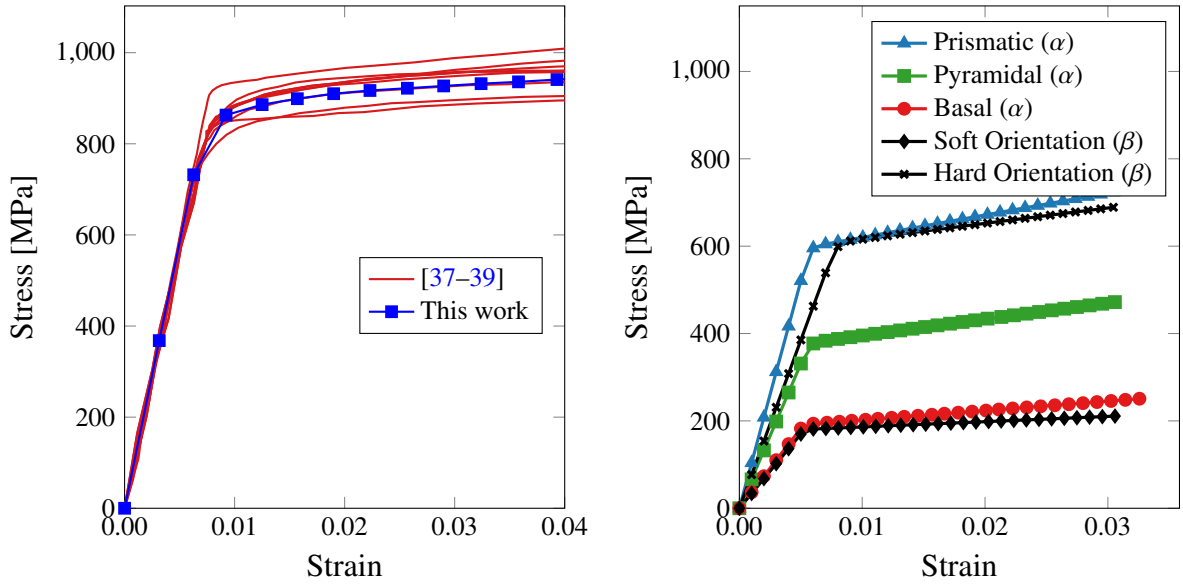
where C is the elasticity matrix. The system of equations provided in eqs. (1) to (10) is solved using a finite element approach in systems containing large numbers of grains, as described in Appendix A. A description of the procedure used to construct the crystals is given in Appendix B. The relevant modeling parameters used throughout the simulations are given in Table 1. The $\dot{\alpha}$, γ_o, m, and b parameters were adopted from ref. [29] and ρ_0 (initial dislocation density) was taken as 10^{12} m^{-2}, which is a reasonable value for the HCP [33] and BCC [27] phases. The k_1 and g_0 parameters were used to fit the model to literature data. The use of a single crystal strength parameter, g_o, was adopted for model simplicity. While the α and β slip systems certainly have different crystal strengths the priority of the exercise was to utilize the CPFE approach to generate adequate data to train a predictive regressor. The result

Parameter	Value	Units
k_1	700	–
g_0	322.2	MPa
b	3×10^{-10}	m
ρ_0	10^{12}	m^{-2}
$\tilde{\alpha}$	0.5	–
m	0.02	–
γ_0	0.001	s^{-1}

Table 1: Simulation parameters used in the finite element simulations.

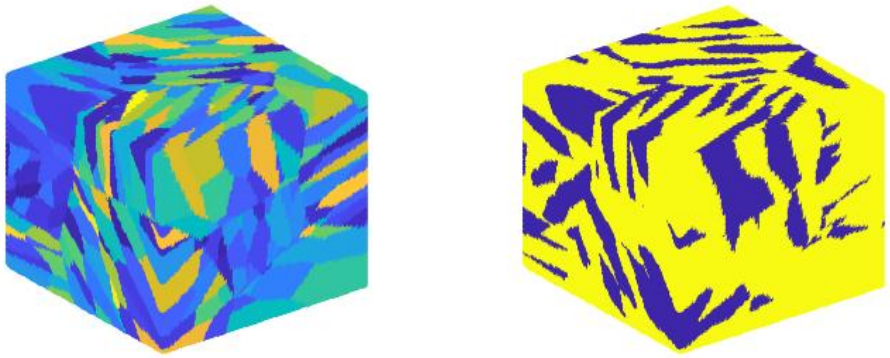
of a simulated tensile test using a 90%/10% α/β equiaxed polycrystal is shown in Figure 1(a) along with a handful of tensile testing results of Ti-6Al-4V from the literature. The results of tensile loading a single crystal of α-Ti in the basal, prismatic, and pyramidal orientations is shown in Fig. 1(b). The anisotropy of the HCP α phase is clearly demonstrated through the varying mechanical response to the different loading orientations. As expected, under basal loading conditions the crystal appears 'soft' while under prismatic loading conditions the crystal appears 'hard'. As well, the strength for two orientations of β-Ti are also shown, a 'soft' one (loading along a direction near the middle of the standard triangle) and a 'hard' one (a vertex of the triangle). As the results show, the α phase can always produce a harder response compared to the BCC β one (partly due to an increased number of available slip systems in the HCP phase, see Appendix A). The β phase has similarly been observed to deform more easily than the α phase in experimental studies [34–36].

A demonstration of the initial configuration of a dual phase crystal is provided in Figure 2 as a function of both phase and texture distribution. For clarity, the initial configuration shown in Fig. 2 is not the configuration that was used to find the data shown in Fig. 1(a).



(a) (b)

Figure 1: (a) Stress-strain curves of a simulated 90/10 α/β polycrystal along with a handful of literature results [37–39]. (b) Simulated tensile testing results from an α-Ti single crystal demonstrating the anisotropic mechanical response of the HCP lattice. Included is a β-Ti single crystal with a random orientation.

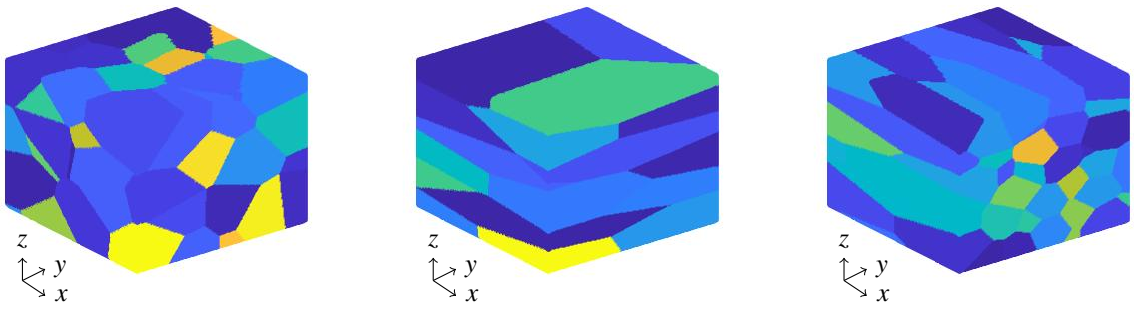


(a) (b)

Figure 2: A 90/10 α/β dual-phase microcrystal in which the coloring represents (a) uniquely oriented lamellae layers and (b) grain composition by color: α (blue) or β (yellow).

For this study we focus on the influence of grain geometry, grain size, strain rate, and α/β volume fraction as contributing factors to the microstructure-controlled strength. Grain geometry refers to the α/β "packets" within the microcrystal. The three grain geometries considered are equiaxed, platelet, and needle, exemplars of which are shown

in Figure 3. Grain size refers to the "packet size" of α/β lamellae. This is an important distinction as each packet is composed of many individual lamellae layers that are approximately 0.5-3 μm thick. For further clarity please refer to the description of the crystal construction in Appendix B. All microstructures were loaded in uniaxial tension along their principal length axis, which was made to coincide with the x direction, as shown in Figure 3.



(a) (b) (c)

Figure 3: Examples of the (a) equiaxed, (b) plate, and (c) needle microstructures. Each color represents a single grain that would contain a packet of uniquely oriented α/β lamellae. All microstructures were placed under uniaxial tension in the x direction.

3. Machine Learning Prediction of Strength and Hardening

Modern crystal plasticity finite element (CPFE) models are continuously improved with the addition of relevant mechanistic information and increased parameter accuracy. These models complement and accelerate experimental efforts by providing indications of pathways to achieve a desired material performance. However, advanced CPFE models are often overly sensitive to certain input parameters and initial conditions and microstructures, making it difficult to parse through extended parameter sets and simulation conditions. In addition, the computational time required to simulate the necessary combinations of unique initial microstructures is not negligible. Therefore, developing a holistic understanding of a material's parametric features is often non-trivial. This effort can be aided by machine learning where regression techniques can be used to generalize the output of CPFE and predict desired micromechanical properties of crystals. Machine learning also provides insight into the importance of different microstructural features to better guide experimental efforts towards the most influential characteristics. In this study we use a supervised machine learning approach to develop several regression models that can assist in the prediction of the mechanical response of a dual-phase titanium polycrystal, namely, the yield strength and hardening rate. Table 2 provides a list of the models utilized in this study and their associated abbreviations. A brief mathematical description of each model is provided in Appendix C. More thorough mathematical descriptions can be found in the references listed in Table 2. Both linear models (e.g., linear regression) and non-linear models (e.g. artificial neural network) were chosen in order to approach the regression exercise with methods that range in complexity. In this way, the simple linear models act as a control condition. Moreover, it has been demonstrated that different regression techniques performer better/worse on different types of data [40–42]. Prior to running a large scale study it is difficult to assess which type of model will perform best, and therefore, we opted to include a variety of regression techniques.

3.1. Evaluation Metrics

Next we define a series of relevant evaluation metrics common to all the regression models:

1. The mean-absolute error (MAE) is the mean of the absolute differences from the predicted, y^i, and true, $\hat{y}$ data from a sampling set of size n. The equation is given as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y^i - \hat{y}\right| \tag{11}$$

Abbreviation	Model	References
LR	Linear Regression	[43]
R-LR	Ridge Linear Regression	[43]
KNN	K-Nearest Neighbors Regression	[43]
RT	Regression Tree	[43]
RF-R	Random Forest Regression	[44, 45]
XGB	`XgBoost`	[40, 46]
GB-R	Gradient Booster Regression	[47–49]
ANN	Artificial Neural Network	[50, 51]

Table 2: List of the models used and their associated abbreviations.

2. The root mean-squared error (RMSE) is the mean square difference from the predicted data and true data. The RMSE is more sensitive to outliers than the MAE and the equation is given as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y^i - \hat{y})^2} \tag{12}$$

3. The R^2 score is a general fit of measure from a predicted regression curve to the original set of data on a scale from 0 (worst) to 1 (perfect). The equation for R^2 is:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y^i - \hat{y}\right)^2}{\sum_{i=1}^{n}\left(y^i - \bar{y}\right)^2} \tag{13}$$

4. Lastly, the Pearson's correlation coefficient (PCC) is used during the results discussion and is described here. The PCC describes the linear correlation between two random variables on a normalized scale between 1 and -1. The measure of 1 being a perfect positive linear correlation and -1 being a perfect negative linear correlation. PCC does not capture non-linear relationships. The equation for PCC is:

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \tag{14}$$

where $\mathbb{E}(x)$ is the expected value of x, μ_X and μ_Y are the means of X and Y, and σ_X and σ_Y their standard deviations. The expectation operator $\mathbb{E}$ here describes the arithmetic mean of the product of individual differences between the random variables (X, Y) and their respective means.

All models were implemented using the `scikit-learn` library and `xgboost` library in `Python` [46, 52]. 8-fold cross-validation was used during training.

4. Results

4.1. Plasticity Model Results

Crystals were constructed using a four-dimensional parameter vector whose components are given in Table 3. The matrix of combinations that results from exploring these four dimensions amounts to 567 unique points in parameter space ($9\times7\times3\times3$ matrix), each of which was run four independent times to ensure numerical consistency and –when relevant– statistical validity. Thus, a total of 2268 finite element simulations were run over the course of this work. Data points were generated at random within the intervals specified in Table 3.

For each simulation, the yield strength was measured by the 0.2%-strain offset rule and the hardening rate was determined as the linear slope of stress-strain curve after yield.

Figure 4 shows the distribution of yield strength and hardening rate as a function of strain rate, β volume fraction, and grain geometry. All subplots contain the same data though different combinations of the input features are used

Parameter	Values	Units
β fraction	5, 10, 15, 20, 25, 30, 35, 40, 45	vol. %
Grain Size	10, 12, 14, 17, 23, 25, 30	μm
Strain Rate	1.0, 5.0, 10.0	10^{-3} s^{-1}
Grain Geometry	Equiaxed, needle, platelet	–

Table 3: Input values for crystal formation used in finite element simulations.

to demonstrate trends in the crystal's mechanical response. In Figure 4(a), the data show that β fraction and strain rate both positively correlate to yield strength, as shown by the tendency for larger red circles on exist the right side of the plot. It is also apparent that the highest hardening rates are typically observed with the lowest strain rate and highest β fraction samples. Figure 4(b) demonstrates that the grain size has a definite positive correlation with the hardening rate for certain samples. The relationship between grain size and yield strength is not discernible from the data shown. As well, grain geometry does not appear to have a clear correlation with either hardening rate or yield strength.

Figure 4(c) reinforces the observations that the lowest β fraction samples exhibit the lowest yield strength and hardening rate while the highest β fraction samples exhibit the highest yield strength and hardening rate. This effect is amplified as the grains get smaller.

Figure 5 shows histograms of hardening rate and yield strength for the all data captured during parametric sweep (data shown in Figs. 4(a) to 4(c). The mean value of each parameter is shown, as well as the value of the first three standard deviations. These markers can be used in coordination with the previous plots to determine the samples that fall more than three standard deviations from the mean and can therefore be considered outliers. The hardening rate data shows a measured mean value of 2.69 GPa with a standard deviation of 1.94 GPa. The yield strength data showed a measured mean value of 883.4 MPa with a standard deviation of 36.5 MPa. Using these criteria, 18 outliers were found, 13 based on hardening rate and 5 based on yield strength. Further examination of the outliers showed each had 4 or fewer grains of 25 μm or larger, 12 had 45% β fraction, 13 had an input strain rate of 10^{-3} s^{-1}.

Figure 6 shows the yield strength or hardening rate as a function of two input parameters. Each row of plots have the same input parameters for ease of comparison. Plots (a) and (b) show the yield strength and hardening rate versus strain rate and β fraction. Plots (c) and (d) show the same outputs as a function of grain size and β fraction. Plots (e) and (f) show the outputs versus grain geometry and grain size. Each plot shows all non-zero data from the simulated data set.

Synthesizing the information shown in Figures 6(a) and 6(b), we see that increasing the β fraction and strain rate together results in increasing yield strengths, but does not contribute to increasing hardening rates. With respect to yield strength, increasing the strain rate from $\dot{\varepsilon} = 0.001$ s^{-1} to $\dot{\varepsilon} = 0.005$ s^{-1} has less effect than changing the strain rate the same amount up to $\dot{\varepsilon} = 0.01$ s^{-1}. In contrast, the largest increase in hardening rate is seen at high β percentage and low strain rate. By the information in these two plots, β percentage alone is not enough to control both the yield strength and post-yield hardening rate. A combination of β percentage and strain rate would be necessary to tune the output yield strength and hardening rate to desired values.

Using the information shown in Figures 6(c) and 6(d) we can further examine the effect of grain size on the output yield strength and hardening rate. According to the output of the simulations, the grain size has little effect on the yield strength until the upper limit of 30 μm was reached. For these specific samples, increasing the β fraction had a stronger positive correlation than in samples with more grains. This trend also exists for the measured hardening rate, as the samples with the fewest grains show a significantly stronger correlation between hardening rate and β fraction. From these two plots it appears grain size has a minimal effect on yield strength and hardening rate at least for a fixed simulation volume (2.7×10^{-14} m^3 in our case).

4.2. Machine Learning Results

Prior to model fitting we investigated the relative importance of the feature variables. Feature importance measurement is similar to traditional forms of sensitivity analysis wherein the emphasis is typically on identifying which features have the most/least impact on the predicted variable. To do this we trained a random forest regressor and calculated the permutation importance for all features. A random forest was chosen to run the preliminary analysis

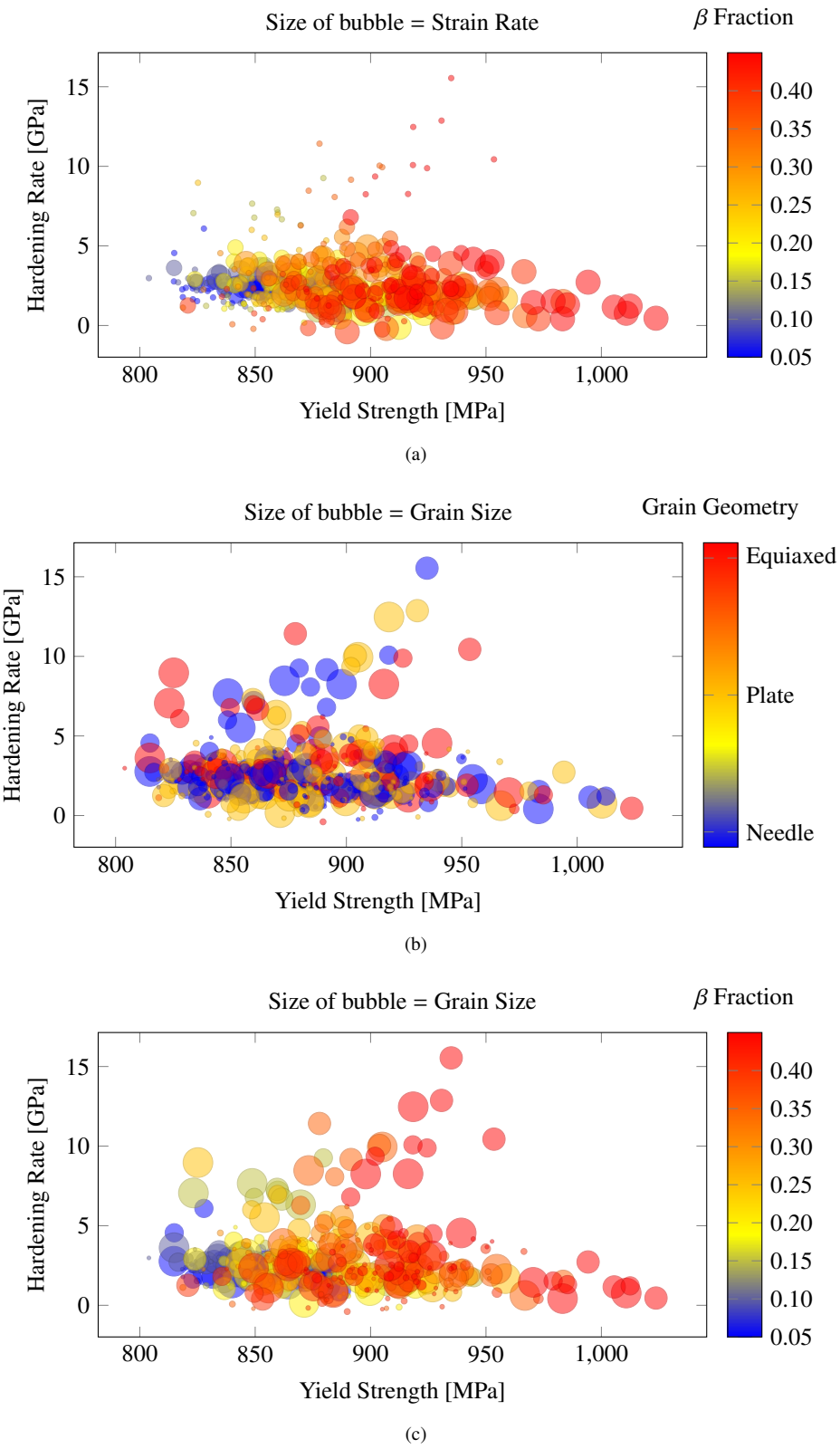


Figure 4: Hardening rate and yield strength distributions for all "gridded" simulations. Same data displayed for each plot with color/bubble size discrimination added for (a) strain rate and β fraction, (b) grain size and grain geometry, and (c) grain size and β fraction.

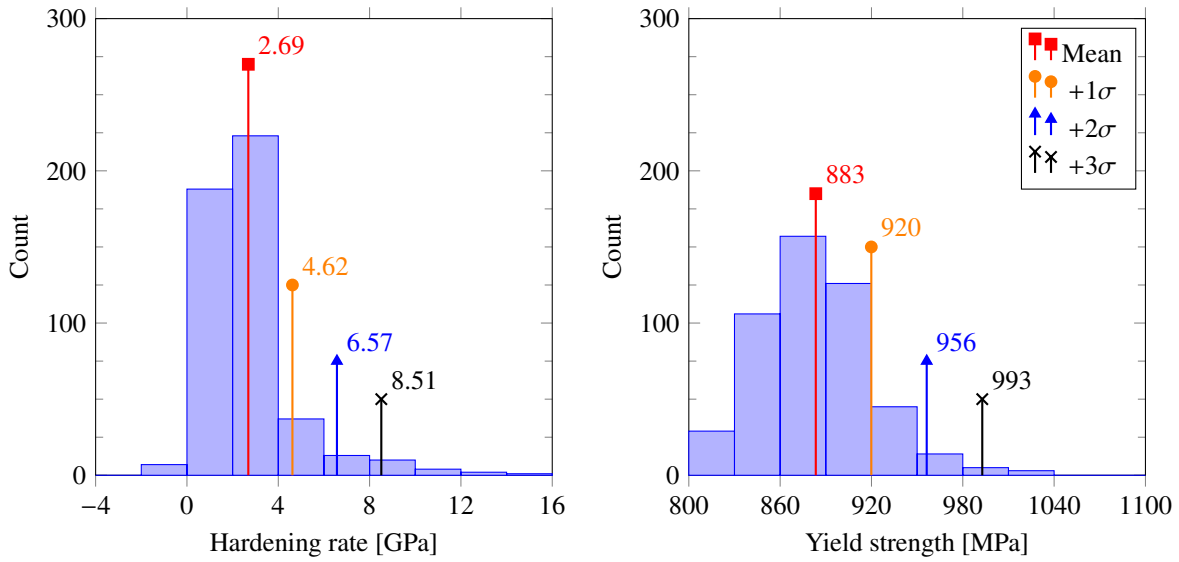


Figure 5: Histograms for data distribution of (a) hardening rate and (b) yield strength. Flags used to show mean and standard deviations of data.

because they have proven to be successful at adapting to a diverse set of problems [53]. The authors refer to the appendix and the attached works for a more thorough description of the regression technique. The permutation importance is calculated by randomly permuting the row-wise order of a given feature and then re-calculating the prediction scoring (RMSE, in this case) of the entire set [54, 55]. This procedure is done individually for all features and the normalized relative importance values are assigned based on the magnitude of decrease in the scoring for the permuted feature whereby the sum of all importance values is 1. The feature with the highest permutation importance decreases the prediction accuracy the greatest (of the feature variables) when its values are randomly permuted. The calculation of permutation importance is a common method to down-select from a high-dimensional feature set [56, 57]. The relative permutation importance values for both the yield strength and hardening rate models are given in Figure 7. For both models the grain geometry is the feature with the least importance (with values of 0.001 and 0.02 for the strength and hardening rates, respectively). All models were trained with the grain geometry feature both included and excluded and the better-performing model was selected. Excluding the grain geometry parameter increased the accuracy of all strength models and raised the average R^2 by 4%. For the hardening rate models the R^2 score increased by an average of 6%.

The skew of the hardening rate and yield strength outputs, 2.56 and 0.60 respectively, were initially considered as detrimental factors to both models' performances. However, re-training with normalized outputs did not meaningfully improve any of the models' behavior. Similarly the 3σ rule was applied to remove outlying data but model performance worsened across the board. During training all continuous features were normalized and the grain geometry feature was one-hot encoded (i.e. equiaxed=1, plate=0, needle=-1). Hyperparameter tuning was performed using 5-fold cross-validation and a grid search method.

The fitting results for the yield strength-predicting ML models is listed in Table 4, while Figure 8 plots the RMSE, MAE, and R^2 results for the testing data as well as the random dataset. All models have R^2 of 0.7 or below. The linear models (LR,R-LR) have performed similarly with R^2 values of 0.59 while the other four models (RFR, RFR, XGB, GB-R,ANN) have R^2 values of 0.64 to 0.70. The KNNR method fell in between the two regions with a R^2 value of 0.60. The poor relative performance of the linear models (LR, R-LR) is expected because the strength-microstructure relationships tend to be nonlinear.

The fitting results for the hardening rate-predicting ML models are given in Table 5, and the associated testing RMSE, MAE, and R^2 values are plotted in Figure 9. Again the non-linear models (KNNR, RT, RFR, XGB, GB-R, ANN) outperform the linear models (LR, R-LR). The hardening rate predictors do not perform as well as the yield

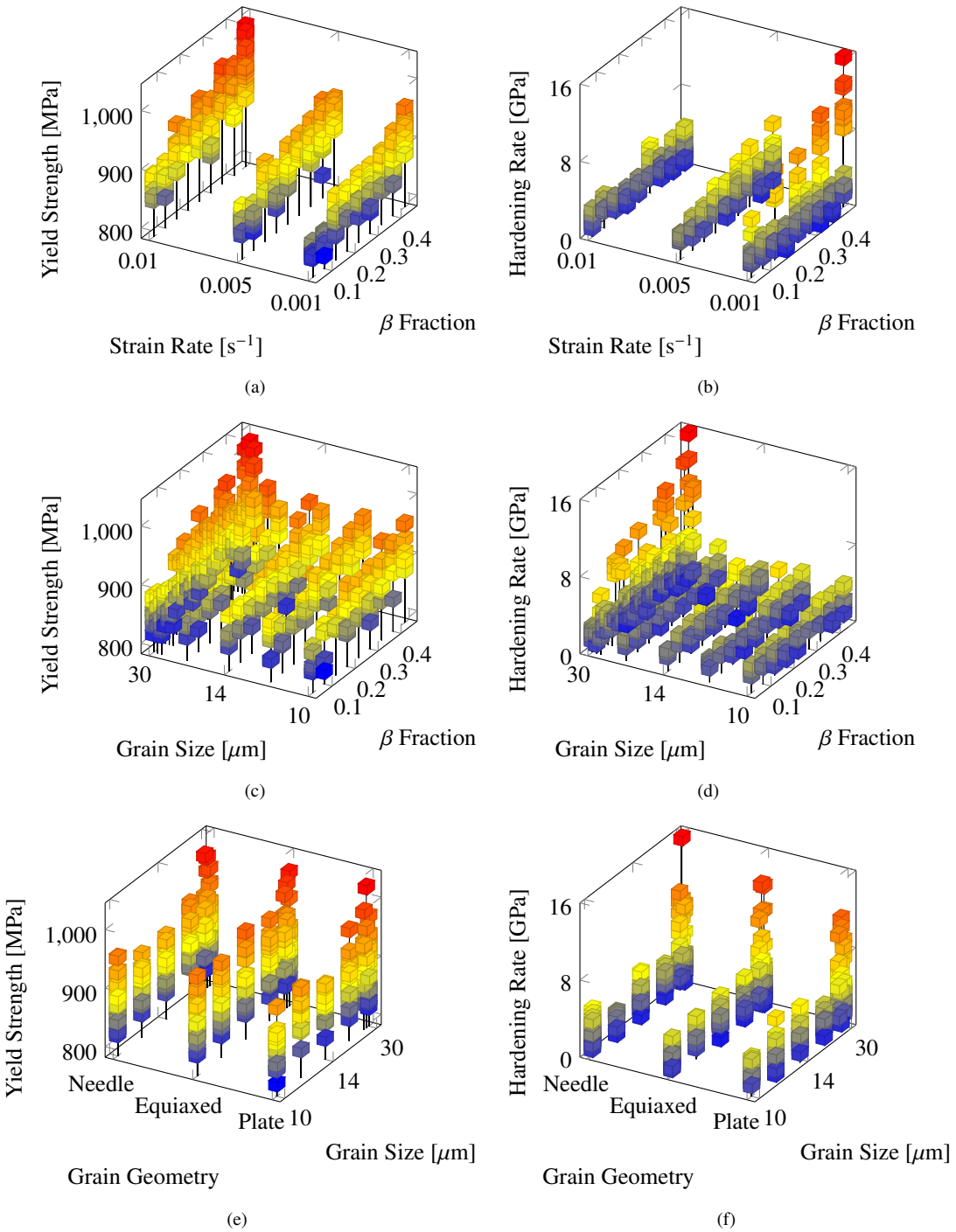


Figure 6: Yield strength and hardening rate plots versus multiple input parameters

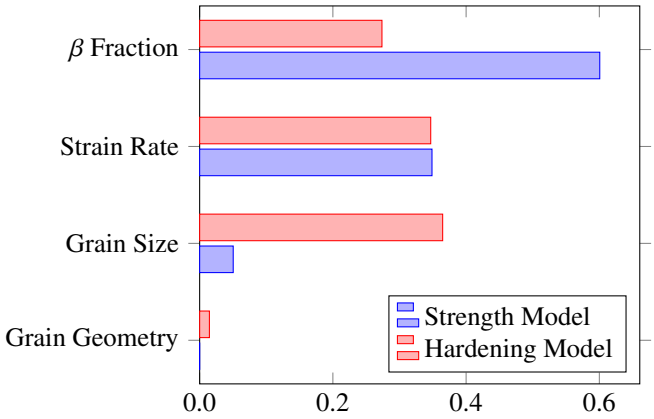


Figure 7: Feature importance as calculated by feature permutation with the trained random forest regression model. The "strength model" and "hardening model" correspond to the preliminary random forest regression models trained on the yield strength and hardening rate data, respectively.

strength predictors and the best hardening rate model (in terms of R^2) is RFR with R^2=0.62.

The distribution of true and predicted values for the the test evaluation of both the regularly-spaced and random datasets for several models are provided in Appendix D.

Model	RMSE Test	MAE Test	R^2	RMSE Random
LR	22.3	17.3	0.59	25.4
R-LR	22.3	17.3	0.59	25.3
KNNR	21.4	16.7	0.62	24.8
RT	20.0	15.7	0.67	29.9
RFR	19.6	15.3	0.68	25.4
XGB	19.2	15.3	0.70	25.3
GB-R	19.3	15.2	0.69	25.6
ANN	20.8	15.6	0.64	24.8

Table 4: Model performance for predicting yield strength values. RMSE and MAE values are units of MPa.

Model	RMSE Test	MAE Test	R^2	RMSE Random
LR	1.78	1.26	0.10	1.08
R-LR	1.78	1.24	0.10	1.07
KNNR	1.13	0.86	0.59	1.34
RT	1.12	0.82	0.60	1.44
RFR	1.10	0.81	0.62	1.38
XGB	1.16	0.85	0.57	1.29
GB-R	1.17	0.87	0.57	1.36
ANN	1.34	0.86	0.47	1.16

Table 5: Model performance for predicting hardening rates. RMSE and MAE values are in units of GPa.

Figure 10 plots the target output of all unique points in the parametric space ordered from least to greatest for yield strength and hardening rate. The noise to signal ratio, $\sigma/\bar{x}$, is overlaid for each plot as red points. The σ values are the standard deviation of the four replica simulations run for each point and the $\bar{x}$ value is the mean that was used for model training. For the yield stress the signal to noise ratio is consistent for all points at values between 0.02 to 0.08. In contrast, the signal to noise ratio for the hardening rate data reaches values as high as 8.0 for the lowest

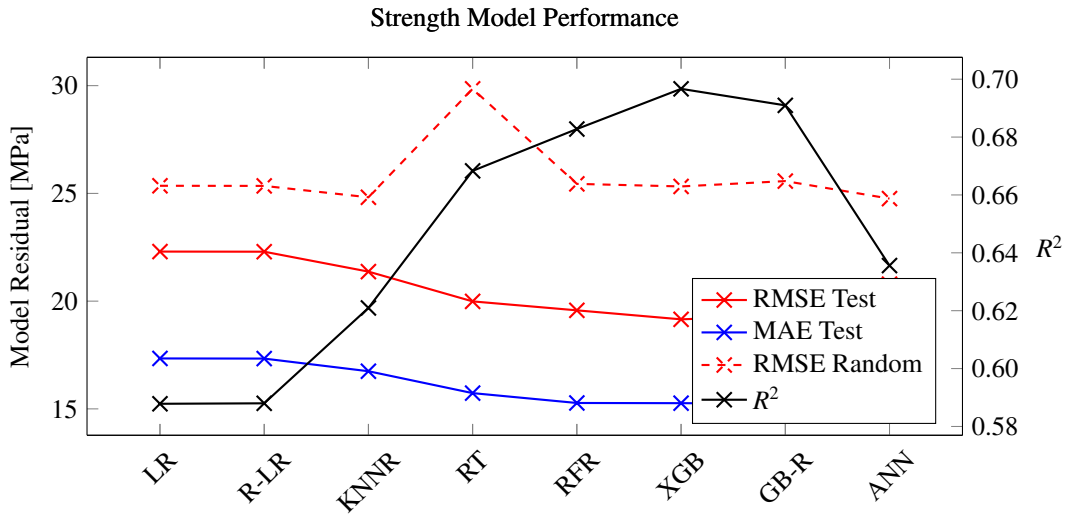


Figure 8: Performance of the models trained for fitting the strength data. Computed using a randomly selected test set of 100 samples.

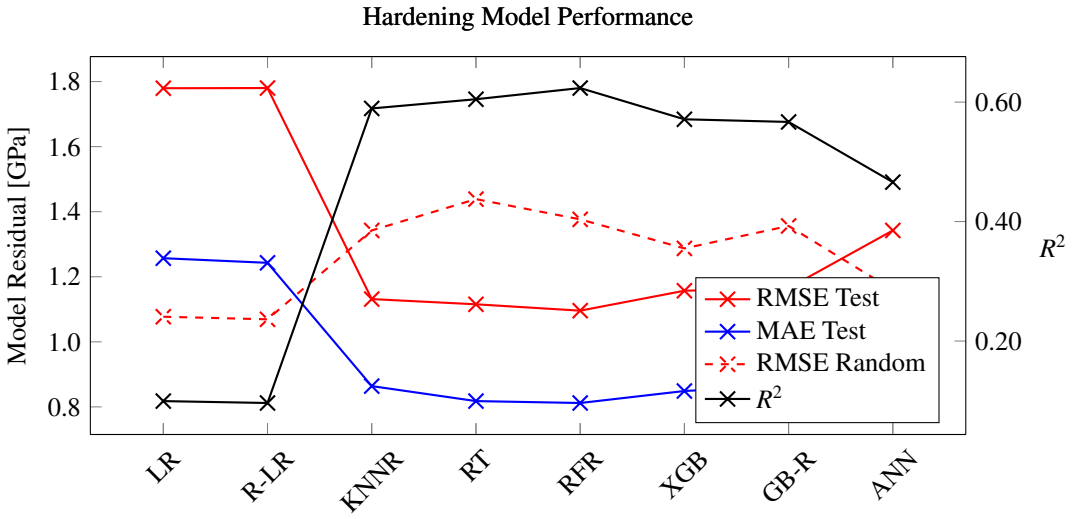
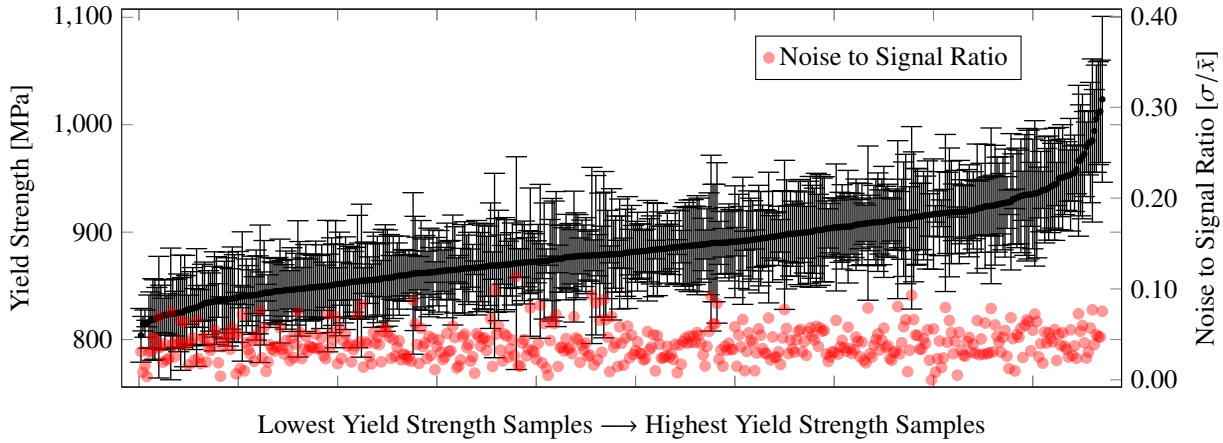
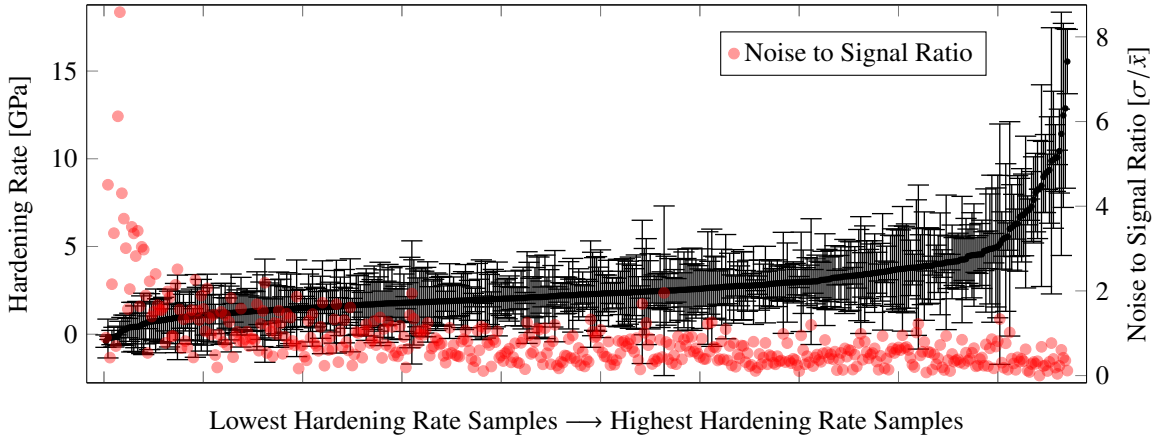


Figure 9: Performance of the models trained for fitting the strength hardening data. Computed using a randomly selected test set of 100 samples.

hardening rate samples and and has an average value of approximately 1.0. The comparatively high $\sigma/\bar{x}$ value for the hardening rate data indicates that the hardening rate replicas had poor agreement with one another. Statistical variation in the target values is not a detriment to the microstructural model. Rather, it is expected that the random sampling construction procedure generates a unique microstructure for each simulation and certain micromechanical responses (e.g., hardening rate) are more sensitive to the stochastic nature of the construction than others. A detailed investigation of these parameters is beyond the scope of this paper and will be investigated in future work.



(a)



(b)

Figure 10: Plot of the mean target value from a unique point in simulation space ordered from least to greatest in terms of (a) yield strength and (b) hardening rate. Error bars are the standard deviation from 4 replicas simulated at each point. The red circles are the noise to signal ratio $\sigma/\bar{x}$.

Further understanding of the performance can be determined with the evolution of pair-wise correlations between the feature variables and the target variables based on the number of simulations conducted. This provides information on how many simulations are needed before the model reaches a maximum in accuracy for each individual correlation value. This also provides information about the change in the correlation values, and how many simulations are needed before reliable information can be obtained from the simulated data set. Figure 11 shows the Pearson correlation coefficient versus the number of unique data points simulated for both yield strength and hardening rate based on the four input parameters. As the figure shows, convergence in the predictors is achieved after 200 samples approximately. The yield strength plot confirms the positive correlation from β fraction and strain rate, and the minimal correlation with grain size and geometry. The hardening rate plot shows the near-zero positive correlation from β fraction and geometry, and the negative correlation from grain size and strain rate.

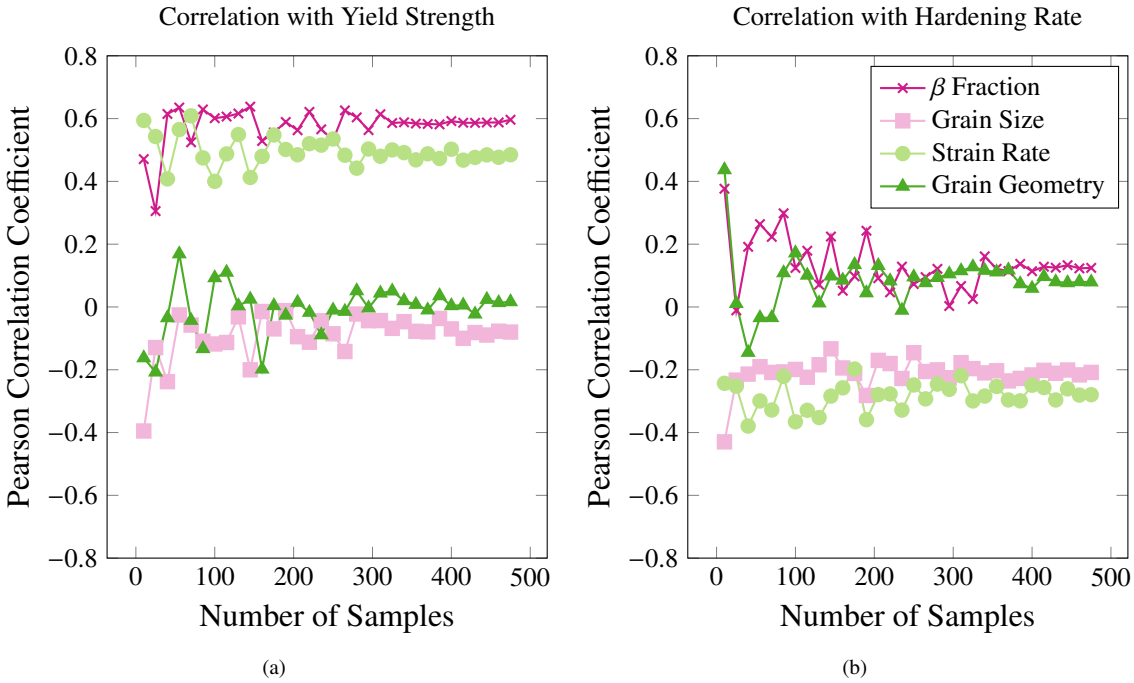


Figure 11: Pearson correlation coefficient versus number of samples for (a) yield strength and (b) hardening rate.

5. Discussion

The main purpose of this study is to demonstrate how a relative large parameter set (defined by four microstructural variables, with several values each, 567 distinct simulation conditions in total) can be parsed effectively using machine learning techniques to predict useful outcomes in terms of alloy mechanical properties. Ultimately, studies such as the present one are aimed at improving and refining the alloy design process to save scarce resources, both in terms of time and money, and accelerate material characterization and synthesis by focusing on parameters with the greatest influence. Specifically, we have chosen a system with a relatively complex microstructure but great metallurgical promise, namely Ti-6Al-4V alloys with various features. Indeed, Ti-6Al-4V has recently been the subject of design optimization efforts using machine learning techniques [58, 59]. Next, we discuss our main findings and identify the lessons learned and their potential applicability.

5.1. Plasticity Model Discussion

The constitutive law and flow rule used in this work have been chosen due to their simplicity so that –in principle– they lead to uncomplicated material responses to facilitate the extraction of trends using the machine learning methods. However, they are still grounded on solid crystal plasticity principles and some interesting results are worth being discussed.

For example, the yield stress displays near-logarithmic growth with strain rate for β fractions up to roughly 30% vol, as shown in Fig. 13(a) This agrees well with previous Ti-6Al-4V studies [38, 39, 60] and accurately reflects the basis of a flow stress power law. When the β fraction is greater than 30%, the σ_y-$\dot{\varepsilon}$ relationship transitions to being exponential. The amount of β in the crystal plays an increasingly important role in determining the magnitude of the yield strength as $\dot{\varepsilon}$ is increased (see Figure 6(b)). The β fraction and σ_y relationship is near linear which agrees with previous works [61] and should be expected as a function of the general rule of mixtures. This is demonstrated in Fig. 13(b). In a similar study of Ti-6Al-4V deformed dynamically (at rates larger than those considered here), the slopes of the β fraction and σ_y relation were observed to increase with strain rate [61]. In the crystal plasticity model employed here, neither phase displays intrinsic hardening (recall that $k_2 = 0$ for both the α and β phases in Table 1).

Thus, the sole source of hardening is given by the value of the resolved shear stress itself, which as a general rule is always higher for systems with a reduced number of slip systems (and the associated lattice stress). As indicated in Table A.6, the HCP α phase contains a total of 39 independent slip systems, against the 12 of the BCC β phase. It is thus reasonable to obtain an increase in the hardening rate as the volume fraction of β increases.

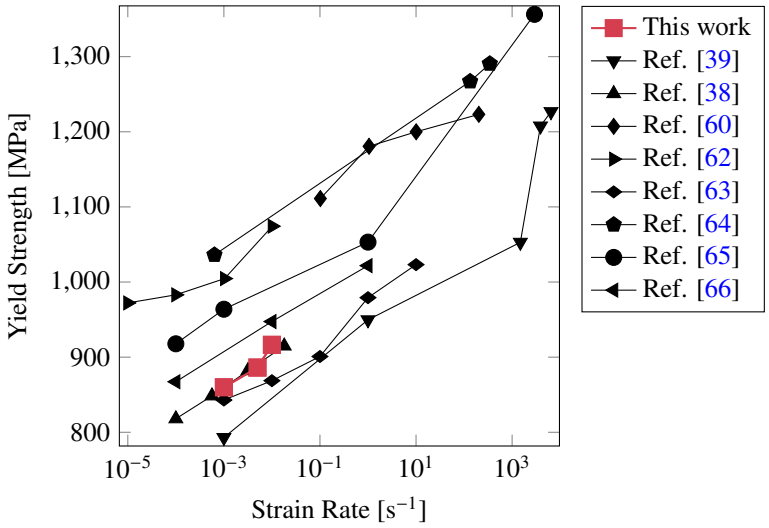


Figure 12: Yield strength as a function of strain rate for this study and experimentally determined values from literature.

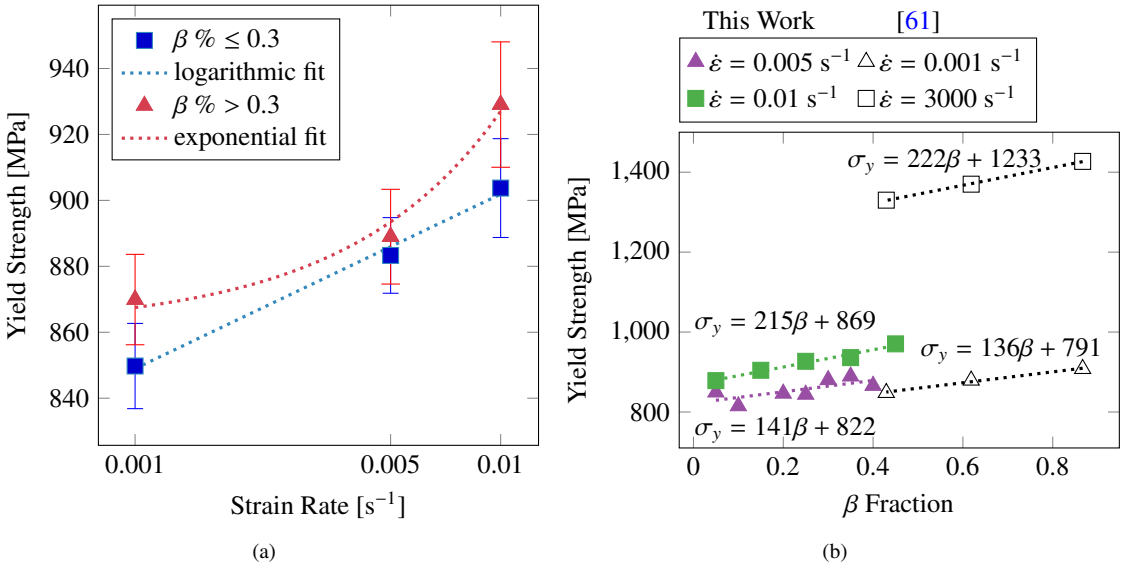


Figure 13: (a) A subset of the data partitioned by crystals with β % ≤ 0.3 (in blue) and β % > 0.3 (in red). The crystals with β % ≤ 0.3 have a logarithmic relationship between σ_y and $\dot{\varepsilon}$. Meanwhile the crystals with β % > 0.3 follow an exponential trend for σ_y and $\dot{\varepsilon}$ for the range explored. (b) Yield strength as a function of β fraction for a subset of $\dot{\varepsilon}$ = 0.005 s^{-1} and $\dot{\varepsilon}$ = 0.01 s^{-1} data plotted alongside data from [61].

Regarding grain geometry, in this study it was seen to have practically no effect on the measured yield strength or hardening rate, as shown in Figs. 6(e) and 6(f). Grain geometry effects may have been minimized due to the choice to simulate lamellae with minimal microstructural texture, as opposed to adding texture as another input parameter and expanding the parameter space. Though each α or β layer had its own unique orientation, there was no difference

in interlayer misorientation between the needle, equiaxed, and plate geometries. Thus each grain geometry resulted in low texture crystals that differed in grain shape but conserved grain boundary area and texture. This negates the anisotropy of the HCP phase, making the simulation volume directionally-independent. This may also explain the presence of outliers in the simulations with larger grain sizes (25-30 μm) as fewer grains may randomly be oriented in harder or softer directions. Though beyond the scope of this work, this effect can be corrected by sampling lamellar packets that are constructed with preferential orientations that consider the grain shape and orientation itself. This type of adjustment could then provide insight to the difference in mechanical response provided by plate, needle, and equiaxed structures with lamellar and non-lamellar substructures that are oriented different ways within the outer grain structures.

The yield strength data produced in this study deviates from the classical Hall-Petch relationship (see Figure 6(c)) when the dislocation mean free path is considered to be the grain size. The limited influence of grain size on the strength and hardening is likely a result of the lamellar substructure. Since the lamellae packing does not differ in terms of spacing or density between large/small grains the total inter-lamellar distance remains roughly constant despite variations in the grain size (for the range explored). This effect could be mitigated by imposing a local hardening condition that is reflective of the dislocation pileup due to the true grain size of local lamellae packet. Alternatively a non-lamellar substructure (i.e., large α and β grains) or Hall-Petch-type strengthening parameter could be used.

Furthermore, is has been experimentally observed that the α/β lamellar thickness and ratio in Ti-6Al-4V strongly dictates mechanical behavior [67–71]. It is possible that negligible influence of grain size may, rather, be due to the α/β lamellae width effects having a dominant influence on strength. Figure 14 shows the yield strength as a function of lamellar spacing and α lamellae width for this study and experimental data found in literature. Note that since we used constant lamellae packet size (i.e., the width of a single α/β bilayer) there is a single data point from this work in Figure 14(a). The Hall-Petch type relationship is recovered for yield strength when the dislocation free glide distance is considered to be the α lamellae width, as shown in Figure 14(b). Considering that BCC β phase is generally softer than the HCP β phase (see Figure 1(b)), increasing the β content would be expected, to first order, to lead to a decrease in strength. However, here we see that the Hall-Petch strengthening achieved through the reduction in the α lamellae widths more than compensates and increases the strength of the alloy. Future work will include refinement of the lamellar spacing/grain size relationship in the context of the dual-phase CPFE model.

Further improvements can be made to the model by incorporating mechanisms such as grain boundary strengthening, reinforcement particle strengthening, precipitation/solid solution strengthening, and temperature effects by modifying the crystal strength and plasticity expressions (equations 6-8). Consideration of non-lamellar microstructures and dynamic loading effects may be beneficial to assist with modeling high-strain rate phenomena.

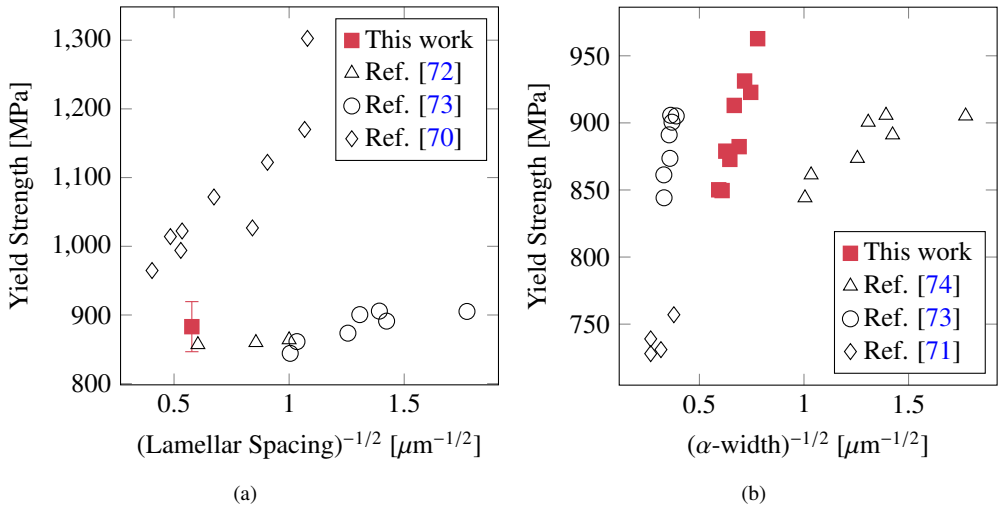


Figure 14: Yield strength as a function of (a) lamellar spacing and (b) α width for this study and experimentally determined values from literature. Lamellar spacing is considered the width of a combined α/β bilayer.

5.2. Machine Learning Discussion

Multidimensional parametric design spaces can be complex to analyze by 'hand' and are further complicated by non-linear feature behavior. To aid with analysis of these types of design spaces the ultimate goal of many machine learning regression exercises is to provide a predictive model for target values learned through example data. In this case, the regression exercise was aimed at modeling the yield strengths and hardening rate of a dual-phase Ti crystal with certain microstructural traits. Several models were trained for both target values using a uniformly-spaced dataset and then evaluated using both the uniform grid and random dataset. An ensemble regressor was constructed from all the trained models [75]. An ensemble regressor makes a prediction based off of the weighted average of individual predictions from several multiple models. In our case, the weight prescribed to each model was calculated using the arbitrary expression:

$$w_i = \frac{1}{3}\left[\left(\frac{\mathrm{RMSE}^i_{\mathrm{test}}}{\sum \mathrm{RMSE}^i_{\mathrm{test}}}\right)^{-1} + 2\left(\frac{\mathrm{RMSE}^i_{\mathrm{random}}}{\sum \mathrm{RMSE}^i_{\mathrm{random}}}\right)^{-1}\right] \tag{15}$$

where $\mathrm{RMSE}^i_{\mathrm{test}}$ and $\mathrm{RMSE}^i_{\mathrm{random}}$ are the testing root-mean-squared errors calculated using the test and random datasets, respectively. Because true microstructural traits exist on a continuum scale, we placed an additional (arbitrary) weight on the performance of the models on the random dataset. Predicted yield strength and hardening rate values as a function of the β fraction and grain size from the voting models are given in Figure 15. Though it is beyond the scope of this study, the authors note that the influence of randomly sampled data on an otherwise uniformly distributed parametric dataset is an interesting premise for improving a model's predictive power. That is, how much "non-grid" data should be added to an otherwise organized dataset in order to achieve an acceptably generalized model.

Lastly, it is important to recognize that the predictive power of the models could undoubtedly be improved with more data. Increasing the number of replicas at each point in parametric space would help further reduce the noise to signal ratio, particularly for the hardening rate data. This is evidenced by Figure 16 which shows the average signal to noise ratio as a function of the number of replicas. 100 structures were selected at random to generate a fifth replica. Similarly the models would benefit from a "finer gridded" parametric space (e.g., simulating structures with $\dot{\epsilon} = 0.0075\ \mathrm{s}^{-1}$, $\dot{\epsilon} = 0.002\ \mathrm{s}^{-1}$) that would simply provide a more rich training set.

6. Conclusions

We conclude the paper with our most important findings:

- We have extended a single-crystal plasticity model to study a polycrystal dual phase material with complex microstructures. The model captures dual-phase BCC/HCP microstructures using standard dislocation evolution models with features inspired in experimental behavior.
- Our model agrees with the tensile testing behavior observed in other works and does well to capture trends in crystal strength.
- Several machine learning regression models were trained on the data to produce ensemble models that can make quick predictions and generalize the yield strength and hardening rate CPFE outputs.
- We demonstrated clear trends in yield strength and hardening rate as a function of β fraction, strain rate, grain geometry, and grain size.
- As general conclusions, (i) the grain shape has practically no bearing on yield strength and hardening rate outcomes, (ii) the β-phase volume fraction was seen to be the most influential feature on both outcomes, (iii) strain rate is a strong predictor of yield strength but not of hardening rate, while grain size is weakly and negatively correlated with yield strength and hardening, respectively.
- Future work will be aimed at extending the plasticity model to include temperature, obstacles, and dynamic loading conditions.

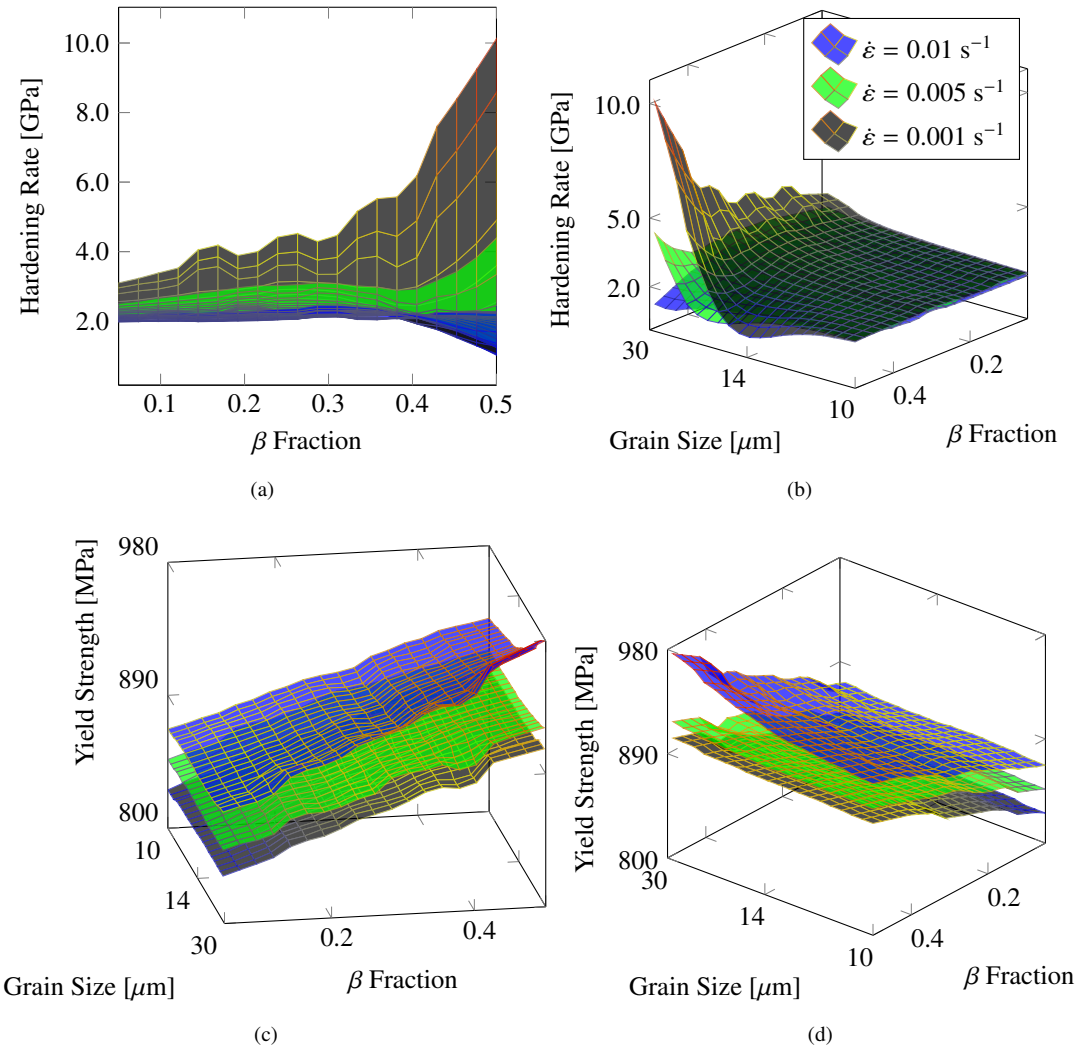


Figure 15: Predictions of (a)-(b) hardening rate and (c)-(d) yield strength as a function of grain size and β fraction for weighted voting regressor. Each plot contains three planes are plotted in black, green, and blue that correspond to $\dot{\epsilon}$ values of 0.001, 0.005, and 0.01 s^{-1}, respectively.

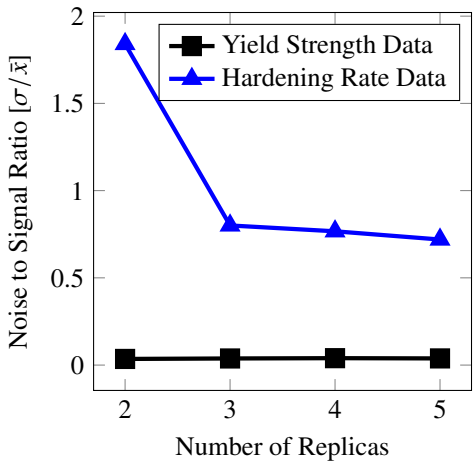


Figure 16: Noise to signal ratio as a function of the number of replicas for the yield strength and hardening rate data

Acknowledgments

We acknowledge support by the National Science Foundation under Grant DMR-1611342, the US Department of Energy's Office of Fusion Energy Sciences, Project DE-SC0012774, and NATO's Science for Peace and Security, Project G5787. Computer time allocations at UCLA's IDRE Hoffman2 supercomputer are acknowledged.

Data availability

The raw/processed data required to reproduce these findings cannot be shared at this time due to technical or time limitations.

References

[1] Lai-Chang Zhang, Liang-Yu Chen, and Liqiang Wang. Surface modification of titanium and titanium alloys: Technologies, developments, and future interests. *Advanced engineering materials*, 22(5), 2020-05.

[2] D. Eylon, S. Fujishiro, P. J. Postans, and F. H. Froes. High-temperature titanium alloys–a review. *J.Met.*, 36(11):55–62, 1984.

[3] Pavlo E Markovsky, Jacek Janiszewski, Vadim I Bondarchuk, Oleksandr O Stasyuk, Dmytro G Savvakin, Mykola A Skoryk, Kamil Cieplak, Piotr Dziewit, and Sergey V Prikhodko. Effect of strain rate on microstructure evolution and mechanical behavior of titanium-based materials. *Metals*, 10(11):1404, 2020.

[4] Gerd Lütjering and James C Williams. *Titanium*. Springer Science & Business Media, 2007.

[5] Lore Thijs, Frederik Verhaeghe, Tom Craeghs, Jan Van Humbeeck, and Jean-Pierre Kruth. A study of the microstructural evolution during selective laser melting of ti–6al–4v. *Acta materialia*, 58(9):3303–3312, 2010.

[6] Sergey V Prikhodko, Pavlo E Markovsky, Dmytro G Savvakin, Oleksandr Stasiuk, and Orest M Ivasishin. Thermo-mechanical treatment of titanium based layered structures fabricated by blended elemental powder metallurgy. In *Materials Science Forum*, volume 941, pages 1384–1390. Trans Tech Publ, 2018.

[7] Yoshiki Kawano, Tetsuya Ohashi, Tsuyoshi Mayama, Masatoshi Mitsuhara, Yelm Okuyama, and Michihiro Sato. Crystal plasticity analysis of microscopic deformation mechanisms and gn dislocation accumulation depending on vanadium content in β phase of two-phase ti alloy. *Materials Transactions*, 60(6):959–968, 2019.

[8] Kartik Kapoor, Priya Ravi, Ryan Noraas, Jun-Sang Park, Vasisht Venkatesh, and Michael D Sangid. Modeling ti–6al–4v using crystal plasticity, calibrated with multi-scale experiments, to understand the effect of the orientation and morphology of the α and β phases on time dependent cyclic loading. *Journal of the Mechanics and Physics of Solids*, 146:104192, 2021.

[9] Tang Bin, Xie Shao, Liu Yi, Han Fengbo, Kou Hongchao, and Li Jinshan. Crystal plasticity finite element study of incompatible deformation behavior in two phase microstructure in near β titanium alloy. *Rare Metal Materials and Engineering*, 44(3):532–537, 2015.

[10] JR Mayeur and DL McDowell. A three-dimensional crystal plasticity model for duplex ti–6al–4v. *International journal of plasticity*, 23(9):1457–1485, 2007.

[11] S Aubry, M Rhee, G Hommes, VV Bulatov, and A Arsenlis. Dislocation dynamics in hexagonal close-packed crystals. *Journal of the Mechanics and Physics of Solids*, 94:105–126, 2016.

[12] Gabriel R Schleder, Antonio CM Padilha, Carlos Mera Acosta, Marcio Costa, and Adalberto Fazzio. From dft to machine learning: recent approaches to materials science–a review. *Journal of Physics: Materials*, 2(3):032001, 2019.

[13] Dominik Steinberger, Hengxu Song, and Stefan Sandfeld. Machine learning-based classification of dislocation microstructures. *Frontiers in Materials*, 6:141, 2019.

[14] Venkatesh Botu and Rampi Ramprasad. Adaptive machine learning framework to accelerate ab initio molecular dynamics. *International Journal of Quantum Chemistry*, 115(16):1074–1083, 2015.

[15] Pinar Acar. Machine learning reinforced crystal plasticity modeling under experimental uncertainty. *AIAA Journal*, 58(8):3569–3576, 2020.

[16] Tim Mueller, Aaron Gilad Kusne, and Rampi Ramprasad. Machine learning in materials science: Recent progress and emerging applications. *Reviews in Computational Chemistry*, 29:186–273, 2016.

[17] Jing Wei, Xuan Chu, Xiang-Yu Sun, Kun Xu, Hui-Xiong Deng, Jigen Chen, Zhongming Wei, and Ming Lei. Machine learning in materials science. *InfoMat*, 1(3):338–358, 2019.

[18] Dane Morgan and Ryan Jacobs. Opportunities and challenges for machine learning in materials science. *Annual Review of Materials Research*, 50:71–103, 2020.

[19] Shun Guo, Jinxin Yu, Xingjun Liu, Cuiping Wang, and Qingshan Jiang. A predicting model for properties of steel using the industrial big data based on machine learning. *Computational Materials Science*, 160:95–104, 2019.

[20] Jie Xiong, TongYi Zhang, and SanQiang Shi. Machine learning of mechanical properties of steels. *Science China Technological Sciences*, 63:1247–1255, 2020.

[21] Tien-Thinh Le. Prediction of tensile strength of polymer carbon nanotube composites using practical machine learning method. *Journal of Composite Materials*, 55(6):787–811, 2021.

[22] Kai Yang, Xinyi Xu, Benjamin Yang, Brian Cook, Herbert Ramos, NM Anoop Krishnan, Morten M Smedskjaer, Christian Hoover, and Mathieu Bauchy. Predicting the young's modulus of silicate glasses using high-throughput molecular dynamics simulations and machine learning. *Scientific reports*, 9(1):1–11, 2019.

[23] Jie Xiong, Tong-Yi Zhang, and San-Qiang Shi. Machine learning prediction of elastic properties and glass-forming ability of bulk metallic glasses. *MRS Communications*, 9(2):576–585, 2019.

[24] Jun Cai, Fuguo Li, Taiying Liu, Bo Chen, and Min He. Constitutive equations for elevated temperature flow stress of ti–6al–4v alloy considering the effect of strain. *Materials & Design*, 32(3):1144–1151, 2011.

[25] Xiaoqiang Li, Guiqiang Guo, Junjie Xiao, Nan Song, and Dongsheng Li. Constitutive modeling and the effects of strain-rate and temperature on the formability of ti–6al–4v alloy sheet. *Materials & Design*, 55:325–334, 2014.

[26] Woei-Shyan Lee and Chi-Feng Lin. Plastic deformation and fracture behaviour of ti–6al–4v alloy loaded with high strain rate under various temperatures. *Materials Science and Engineering: A*, 241(1-2):48–59, 1998.

[27] Nikhil Chandra Admal, Giacomo Po, and Jaime Marian. Diffuse-interface polycrystal plasticity: expressing grain boundaries as geometrically necessary dislocations. *Materials Theory*, 1(1):1–16, 2017.

[28] Nathan R Barton, Athanasios Arsenlis, and Jaime Marian. A polycrystal plasticity model of strain localization in irradiated iron. *Journal of the Mechanics and Physics of Solids*, 61(2):341–351, 2013.

[29] John A Moore, Nathan R Barton, Jeff Florando, Rupalee Mulay, and Mukul Kumar. Crystal plasticity modeling of β phase deformation in ti-6al-4v. *Modelling and Simulation in Materials Science and Engineering*, 25(7):075007, 2017.

[30] JJ Fundenberger, MJ Philippe, F Wagner, and C Esling. Modelling and prediction of mechanical properties for materials with hexagonal symmetry (zinc, titanium and zirconium alloys). *Acta materialia*, 45(10):4041–4055, 1997.

[31] SL Semiatin and TR Bieler. Effect of texture and slip mode on the anisotropy of plastic flow and flow softening during hot working of ti-6al-4v. *Metallurgical and Materials Transactions A*, 32(7):1787–1799, 2001.

[32] NE Paton. The deformation of α-phase titanium. *Titanium science and technology*, 1973.

[33] Bijish Babu and Lars-Erik Lindgren. Dislocation density based model for plastic deformation and globularization of ti-6al-4v. *International Journal of Plasticity*, 50:94–108, 2013.

[34] Yan Chong, Tilak Bhattacharjee, Myeong-Heom Park, Akinobu Shibata, and Nobuhiro Tsuji. Factors determining room temperature mechanical properties of bimodal microstructures in ti-6al-4v alloy. *Materials Science and Engineering: A*, 730:217–222, 2018.

[35] Yukimi Tanaka, Koichiro Hattori, and Yoshihisa Harada. Micro-cantilever testing of microstructural effects on plastic behavior of ti–6al–4v alloy. *Materials Science and Engineering: A*, 823:141747, 2021.

[36] S Hémery, P Villechaise, and D Banerjee. Microplasticity at room temperature in α/β titanium alloys. *Metallurgical and Materials Transactions A*, 51(10):4931–4969, 2020.

[37] Adam M. Stapleton, Seema L. Raghunathan, Ioannis Bantounas, Howard J. Stone, Trevor C. Lindley, and David Dye. Evolution of lattice strain in ti-6al-4v during tensile loading at room temperature. *Acta Materialia*, 56(20):6186–6196, 2008.

[38] B.D. Venkatesh, D.L. Chen, and S.D. Bhole. Effect of heat treatment on mechanical properties of ti–6al–4v eli alloy. *Materials Science and Engineering: A*, 506(1):117–124, 2009.

[39] Hongzhi Hu, Zejian Xu, Wang Dou, and Fenglei Huang. Effects of strain rate and stress state on mechanical properties of ti-6al-4v alloy. *International Journal of Impact Engineering*, 145:1, 11 2020.

[40] Kai Song, Feng Yan, Ting Ding, Liang Gao, and Songbao Lu. A steel property optimization model based on the xgboost algorithm and improved pso. *Computational Materials Science*, 174:109472, 2020.

[41] Suraj Kumar Bhagat, Tiyasha Tiyasha, Tran Minh Tung, Reham R Mostafa, and Zaher Mundher Yaseen. Manganese (mn) removal prediction using extreme gradient model. *Ecotoxicology and Environmental Safety*, 204:111059, 2020.

[42] Guibin Dong, Xiucheng Li, Jingxiao Zhao, Shuai Su, RDK Misra, Ruoxiu Xiao, and Chengjia Shang. Machine learning guided methods in building chemical composition-hardenability model for wear-resistant steel. *Materials Today Communications*, 24:101332, 2020.

[43] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[44] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[45] Vladimir Svetnik, Andy Liaw, Christopher Tong, J Christopher Culberson, Robert P Sheridan, and Bradley P Feuston. Random forest: a classification and regression tool for compound classification and qsar modeling. *Journal of chemical information and computer sciences*, 43(6):1947–1958, 2003.

[46] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[47] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[48] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2):337–407, 2000.

[49] Jerome H Friedman. Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4):367–378, 2002.

[50] Jacek M Zurada. *Introduction to artificial neural systems*, volume 8. West St. Paul, 1992.

[51] Bhadeshia Hkdh. Neural networks in materials science. *ISIJ international*, 39(10):966–979, 1999.

[52] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[53] Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168, 2006.

[54] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8(1):1–21, 2007.

[55] Kevin Kaufmann, Daniel Maryanovsky, William M Mellor, Chaoyi Zhu, Alexander S Rosengarten, Tyler J Harrington, Corey Oses, Cormac Toher, Stefano Curtarolo, and Kenneth S Vecchio. Discovery of high-entropy ceramics via machine learning. *Npj Computational Materials*, 6(1):1–9, 2020.

[56] Ankita Mangal and Elizabeth A Holm. A comparative study of feature selection methods for stress hotspot classification in materials. *Integrating Materials and Manufacturing Innovation*, 7(3):87–95, 2018.

[57] Hubert Anysz, Łukasz Brzozowski, Wojciech Kretowicz, and Piotr Narloch. Feature importance of stabilised rammed earth components affecting the compressive strength calculated with explainable artificial intelligence tools. *Materials*, 13(10):2317, 2020.

[58] Chun-Te Wu, Hsiao-Tzu Chang, Chien-Yu Wu, Shi-Wei Chen, Sih-Ying Huang, Mingxin Huang, Yeong-Tsuen Pan, Peta Bradbury, Joshua Chou, and Hung-Wei Yen. Machine learning recommends affordable new ti alloy with bone-like modulus. *Materials Today*, 34:41–50, 2020.

[59] Chengpeng Zhu, Chao Li, Di Wu, Wan Ye, Shuangxi Shi, Hui Ming, Xiaoyong Zhang, and Kechao Zhou. A titanium alloys design method based on high-throughput experiments and machine learning. *Journal of Materials Research and Technology*, 11:2336–2353, 2021.

[60] AJ Wagoner Johnson, CW Bull, KS Kumar, and CL Briant. The influence of microstructure and strain rate on the compressive deformation behavior of ti-6al-4v. *Metallurgical and Materials Transactions A*, 34(2):295–306, 2003.

[61] Yu Ren, Shimeng Zhou, Wenbo Luo, Zhiyong Xue, and Yajing Zhang. Influence of primary α-phase volume fraction on the mechanical properties of ti-6al-4v alloy at different strain rates and temperatures. In *IOP Conference Series: Materials Science and Engineering*, volume 322, page 022022. IOP Publishing, 2018.

[62] SQ Wang, JH Liu, and DL Chen. Effect of strain rate and temperature on strain hardening behavior of a dissimilar joint between ti–6al–4v and ti17 alloys. *Materials & Design (1980-2015)*, 56:174–184, 2014.

[63] Chan Hee Park, Young Il Son, and Chong Soo Lee. Constitutive analysis of compressive deformation behavior of eli-grade ti–6al–4v with different microstructures. *Journal of Materials Science*, 47(7):3115–3124, 2012.

[64] PS Follansbee and GT Gray. An analysis of the low temperature, low and high strain-rate deformation of ti- 6al- 4v. *Metallurgical Transactions*

A, 20(5):863–874, 1989.
[65] A Tabei, FH Abed, GZ Voyiadjis, and H Garmestani. Constitutive modeling of ti-6al-4v at a wide range of temperatures and strain rates. *European Journal of Mechanics-A/Solids*, 63:128–135, 2017.
[66] Akhtar S Khan, Shaojuan Yu, and Haowen Liu. Deformation induced anisotropic responses of ti–6al–4v alloy part ii: A strain rate and temperature dependent anisotropic yield criterion. *International Journal of Plasticity*, 38:14–26, 2012.
[67] Ren Guo Guan, Young Tae Je, Zhan Yong Zhao, and Chong Soo Lee. Effect of microstructure on deformation behavior of ti–6al–4v alloy during compressing process. *Materials & Design (1980-2015)*, 36:796–803, 2012.
[68] Y Ren, SM Zhou, ZY Xue, WB Luo, YJ Ren, and YJ Zhang. Effect of α-platelet thickness on the mechanical properties of ti-6al-4v alloy with lamellar microstructure. In *IOP Conference Series: Materials Science and Engineering*, volume 281, page 012024. IOP Publishing, 2017.
[69] GQ Wu, CL Shi, W Sha, AX Sha, and HR Jiang. Effect of microstructure on the fatigue properties of ti–6al–4v titanium alloys. *Materials & Design*, 46:668–674, 2013.
[70] Xiang-Yu Zhang, Gang Fang, Sander Leeflang, Amarante J Böttger, Amir A Zadpoor, and Jie Zhou. Effect of subtransus heat treatment on the microstructure and mechanical properties of additively manufactured ti-6al-4v alloy. *Journal of Alloys and Compounds*, 735:1562–1575, 2018.
[71] GC Obasi, OM Ferri, T Ebel, and R Bormann. Influence of processing parameters on mechanical properties of ti–6al–4v alloy fabricated by mim. *Materials Science and Engineering: A*, 527(16-17):3929–3935, 2010.
[72] Dong-Geun Lee, Sunghak Lee, Chong Soo Lee, and Sunmoo Hur. Effects of microstructural factors on quasi-static and dynamic deformation behaviors of ti-6al-4v alloys with widmanstätten structures. *Metallurgical and Materials Transactions A*, 34(11):2541–2548, 2003.
[73] ON Senkov, JJ Valencia, SV Senkova, M Cavusoglu, and FH Froes. Effect of cooling rate on microstructure of ti-6al-4v forging. *Materials science and technology*, 18(12):1471–1478, 2002.
[74] Indrani Sen, S Tamirisakandala, DB Miracle, and U Ramamurty. Microstructural effects on the mechanical behavior of b-modified ti–6al–4v alloys. *Acta Materialia*, 55(15):4983–4993, 2007.
[75] Peter Bühlmann. Bagging, boosting and ensemble methods. In *Handbook of computational statistics*, pages 985–1022. Springer, 2012.
[76] Cassie Marker, Shun-Li Shang, Ji-Cheng Zhao, and Zi-Kui Liu. Effects of alloying elements on the elastic properties of bcc ti-x alloys from first-principles calculations. *Computational Materials Science*, 142:215–226, 2018.
[77] Hassel Ledbetter, Hirotsugu Ogi, Satoshi Kai, Sudook Kim, and Masahiko Hirao. Elastic constants of body-centered-cubic titanium monocrystals. *Journal of Applied Physics*, 95(9):4642–4644, 2004.
[78] Desmond Tromans. Elastic anisotropy of hcp metal crystals and polycrystals. *Int. J. Res. Rev. Appl. Sci.*, 6, 01 2011.
[79] F. C. Frank. On Miller–Bravais indices and four-dimensional vectors. *Acta Crystallographica*, 18(5):862–866, May 1965.
[80] David Cereceda, Martin Diehl, Franz Roters, Dierk Raabe, J Manuel Perlado, and Jaime Marian. Unraveling the temperature dependence of the yield strength in single-crystal tungsten using atomistically-informed crystal plasticity calculations. *International Journal of Plasticity*, 78:242–265, 2016.
[81] Qianran Yu, Enrique Martínez, Javier Segurado, and Jaime Marian. A stochastic solver based on the residence time algorithm for crystal plasticity models. *Computational Mechanics*, pages 1–16, 2021.

Appendix A. Finite Element Implementation

The finite element implementation in `COMSOL5.5` for this study is fully described in [27]. A MUMPS direct solver and BDF (Backward Differential Formula) time stepping algorithm were employed. All simulations were performed with a free tetrahedral mesh with 99883 elements and 595620 degrees of freedom. The stiffness matrices for the α and β phases have been taken from refs. [76–78] and are given below (all values in GPa).

$$C^{\alpha} = \begin{bmatrix} 169.4 & 90.0 & 66.0 & 0 & 0 & 0 \\ 90.0 & 169.4 & 66.0 & 0 & 0 & 0 \\ 66.0 & 66.0 & 169.2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 7.4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 7.4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 46.8 \end{bmatrix}$$

$$C^{\beta} = \begin{bmatrix} 119.4 & 55.7 & 55.7 & 0 & 0 & 0 \\ 55.7 & 119.4 & 55.7 & 0 & 0 & 0 \\ 55.7 & 55.7 & 119.4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 31.9 & 0 & 0 \\ 0 & 0 & 0 & 0 & 31.9 & 0 \\ 0 & 0 & 0 & 0 & 0 & 31.9 \end{bmatrix}$$

The CP model admits slip on basal, prismatic and pyramidal planes for the α phase (3, 11, and 25 slip systems respectively), and close-packed slip for the β phase (12 slip systems). All vectors $\boldsymbol{s}$ and $\boldsymbol{m}$ in eq. (3) for the hexagonal

phase are expressed in three-component Miller notation using the conversion introduced by Frank [79]:

$$[h\ k\ i\ l] \quad \rightarrow \quad [(h-i)\ (k-i)\ l] \tag{A.1}$$

$$(h\ k\ i\ l) \quad \rightarrow \quad (h\ k\ l) \tag{A.2}$$

These expressions satisfy the orthogonality relations between slip direction and slip plane normal. The slip systems considered in this work are given in Table A.6. The slip systems for BCC crystals capture $^1/_2\langle 111\rangle\{110\}$ and are given

Table A.6: Conversion from 4-index slip system to 3-index notation

slip system type	No.	slip plane		slip direction	
		4-index	3-index	4-index	3-index
basal	1	(0001)	(001)	$[2\bar{1}\bar{1}0]$	[100]
	2	(0001)	(001)	$[\bar{1}\bar{1}20]$	$[\bar{1}\bar{1}0]$
	3	(0001)	(001)	$[\bar{1}2\bar{1}0]$	[010]
prismatic	4	$(01\bar{1}0)$	(010)	$[2\bar{1}\bar{1}0]$	[100]
	5	$(1\bar{1}00)$	$(1\bar{1}0)$	$[\bar{1}\bar{1}20]$	$[\bar{1}\bar{1}0]$
	6	$(\bar{1}010)$	$(\bar{1}00)$	$[\bar{1}2\bar{1}0]$	[010]
	7	$(01\bar{1}0)$	(010)	$[2\bar{1}\bar{1}3]$	[101]
	8	$(01\bar{1}0)$	(010)	$[2\bar{1}\bar{1}\bar{3}]$	$[10\bar{1}]$
	9	$(1\bar{1}00)$	$(1\bar{1}0)$	$[\bar{1}\bar{1}23]$	$[\bar{1}\bar{1}1]$
	10	$(1\bar{1}00)$	$(1\bar{1}0)$	$[\bar{1}\bar{1}2\bar{3}]$	$[\bar{1}\bar{1}\bar{1}]$
	11	$(\bar{1}010)$	$(\bar{1}00)$	$[\bar{1}2\bar{1}3]$	[011]
	12	$(01\bar{1}0)$	(010)	[0001]	[001]
	13	$(\bar{1}010)$	$(\bar{1}00)$	[0001]	[001]
	14	$(1\bar{1}00)$	$(1\bar{1}0)$	[0001]	[001]
pyramidal	15	$(01\bar{1}1)$	(011)	$[2\bar{1}\bar{1}0]$	[100]
	16	$(0\bar{1}11)$	$(0\bar{1}0)$	$[2\bar{1}\bar{1}0]$	[100]
	17	$(1\bar{1}01)$	$(1\bar{1}1)$	$[\bar{1}\bar{1}20]$	$[\bar{1}\bar{1}0]$
	18	$(\bar{1}101)$	$(\bar{1}11)$	$[\bar{1}\bar{1}20]$	$[\bar{1}\bar{1}0]$
	19	$(\bar{1}011)$	$(\bar{1}01)$	$[\bar{1}2\bar{1}0]$	[010]
	20	$(10\bar{1}1)$	(101)	$[\bar{1}2\bar{1}0]$	[010]
	21	$(\bar{1}011)$	$(\bar{1}01)$	$[2\bar{1}\bar{1}3]$	[101]
	22	$(\bar{1}101)$	$(\bar{1}11)$	$[2\bar{1}\bar{1}3]$	[101]
	23	$(\bar{2}112)$	$(\bar{2}12)$	$[2\bar{1}\bar{1}3]$	[101]
	24	$(10\bar{1}1)$	(101)	$[2\bar{1}\bar{1}\bar{3}]$	$[10\bar{1}]$
	25	$(1\bar{1}01)$	$(1\bar{1}1)$	$[2\bar{1}\bar{1}\bar{3}]$	$[10\bar{1}]$
	26	$(2\bar{1}\bar{1}2)$	$(2\bar{1}2)$	$[2\bar{1}\bar{1}\bar{3}]$	$[10\bar{1}]$
	27	$(10\bar{1}1)$	(101)	$[\bar{1}\bar{1}23]$	$[\bar{1}\bar{1}1]$
	28	$(01\bar{1}1)$	(011)	$[\bar{1}\bar{1}23]$	$[\bar{1}\bar{1}1]$
	29	$(11\bar{2}2)$	(112)	$[\bar{1}\bar{1}23]$	$[\bar{1}\bar{1}1]$
	30	$(\bar{1}011)$	$(\bar{1}01)$	$[\bar{1}\bar{1}2\bar{3}]$	$[\bar{1}\bar{1}\bar{1}]$
	31	$(0\bar{1}11)$	$(0\bar{1}1)$	$[\bar{1}\bar{1}2\bar{3}]$	$[\bar{1}\bar{1}\bar{1}]$
	32	$(\bar{1}\bar{1}22)$	$(\bar{1}\bar{1}2)$	$[\bar{1}\bar{1}2\bar{3}]$	$[\bar{1}\bar{1}\bar{1}]$
	33	$(1\bar{1}01)$	$(1\bar{1}1)$	$[\bar{1}2\bar{1}3]$	[011]
	34	$(0\bar{1}11)$	$(0\bar{1}1)$	$[\bar{1}2\bar{1}3]$	[011]
	35	$(1\bar{2}12)$	$(1\bar{2}2)$	$[\bar{1}2\bar{1}3]$	[011]
	36	$(\bar{1}101)$	$(\bar{1}11)$	$[\bar{1}2\bar{1}\bar{3}]$	$[01\bar{1}]$
	37	$(01\bar{1}1)$	(011)	$[\bar{1}2\bar{1}\bar{3}]$	$[01\bar{1}]$
	38	$(\bar{1}101)$	$(\bar{1}11)$	$[\bar{1}2\bar{1}\bar{3}]$	$[01\bar{1}]$
	39	$(\bar{1}2\bar{1}2)$	$(\bar{1}22)$	$[\bar{1}2\bar{1}\bar{3}]$	$[01\bar{1}]$

in past publications by our group [80, 81].

Appendix B. Construction of Dual Phase Lamellar Polycrystals

All grains were constructed as Voronoi tessellations wherein grain centers c_j were randomly selected from a cubic grid and each individual point p_i was assigned to the nearest grain center such that:

$$p_i^g = \min\left(\text{dist}(p_i, c_j) : j \in \{1, ..., C\}\right) \tag{B.1}$$

where p_i^g is the grain assignment for the point p_i. Here there are a total of C grain centers and the function $dist(p_i, c_j)$ returns the distance between the grid point p_i and the grain center c_j. The distance equation is a modified euclidean distance function give as:

$$\text{dist}(p_i, c_j) = \left(\frac{p_i^x - c_j^x}{sx}\right)^2 + \left(\frac{p_i^y - c_j^y}{sy}\right)^2 + \left(\frac{p_i^z - c_j^z}{sz}\right)^2 \tag{B.2}$$

where sx, sy, and sz are distance scaling factors that enable elongated grains. For this study the scaling factors for the equiaxed, plate, and needle grains are given in Table B.7. To achieve a unique lamellar structure within each grain

Grain Geometry	sx	sy	sz
Equiaxed	1	1	1
Needle	1	1	4
Plate	1	4	4

Table B.7: Grain geometry parameters.

a set of plate-like grains are first constructed. Grains are assigned to either α or β phase as a function of their the distance from the x-axis such that:

$$c(j) = \begin{cases} \beta \text{ phase}, & \text{if } F_\beta/G \leq \min\left(\left|L_j - nl\right| : n \in \{0, ..., G+1\} \text{ where } G = \lfloor H/l \rfloor\right) \\ \alpha \text{ phase}, & \text{otherwise} \end{cases} \tag{B.3}$$

here L is the x coordinate of the grain center, l is the spacing between same-phase lamellae, F_β is the β phase fraction out of 100, and H is the total height of the simulation cube. The $G = \lfloor H/l \rfloor$ term indicates the number of lamellae layers in each simulated cube while the F_β/G term is the thickness of each β layer. Once a plate-like cube was constructed it was put through a set random rotations across the x, y, and z axis to achieve a unique grain alignment. For each simulation a set of 20 unique lamellar crystals were constructed and the sampled from to populate the grain geometry defined by the original Voronoi tesselations –ultimately leading to equiaxed, needle, or plate-like grain geometries with an intra-grain α/β lamellar structure.

We recognize that there are various forms of α/β subgrain morphologies but here we focus on lamellar structures for the purpose of this study.

Appendix C. Machine Learning Regression Models

Appendix C.1. Linear Regression

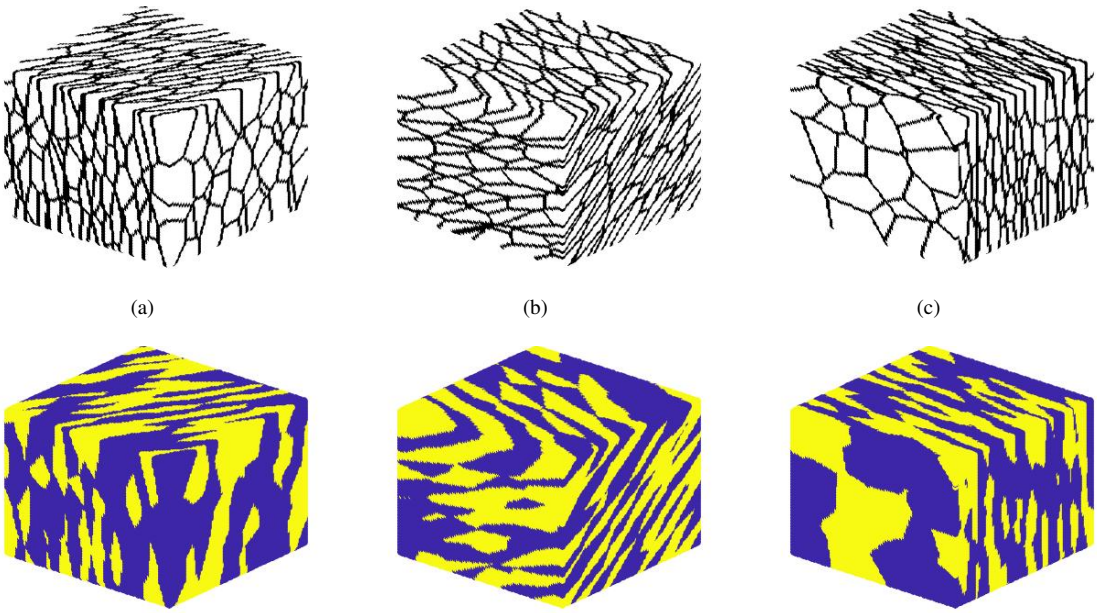
Given the simple hypothesis function of:

$$f(\theta) = \theta_0 x_0 + \theta_1 x_1 + ...\theta_n x_n \tag{C.1}$$

where θ are the feature weights and x_n are the feature values for n features. Considering optimizing a least-squares cost function, a direct solution for the optimal θ values can be expressed as:

$$\theta_{LR} = \left(X^T X\right)^{-1} X^T y \tag{C.2}$$

where X is the collection of input features and y is the output values for each instance. The above expression does not hold for other cost functions, but rather, demonstrates the form of the model.



(a) (b) (c)

(d) (e) (f)

Figure B.17: Pure lamellar crystals with 17(a) - 17(c) outlined phase boundaries and 17(d) - 17(f) shading indicating either α (blue) or β (yellow) phase. All crystals have been put through 3 random rotations across their x, y, and z axes.

Appendix C.2. Ridge Linear Regression

To regularize the traditional linear regression expression in equation C.2 a complexity penalty of $\lambda \sum_{i=1}^{n} \theta_j^2$ can be added to the least-squares cost function. Because the cost function is still convex, there is a unique solution:

$$\theta_{RR} = \left(X^T X + \lambda I\right)^{-1} X^T y \tag{C.3}$$

where $\lambda \in [0, \infty)$ is the regularization parameter and I is the identity matrix.

Appendix C.3. K-Nearest Neighbors Regression

K-nearest neighbors is a non-parametric method that that approximates the value of a new instance by averaging the target value of observations in the same neighborhood. Given a new instance, an estimate is calculated by first finding the K-nearest neighbors in Euclidean space and then averaging the feature set such that:

$$y_i = \frac{1}{k} \sum_{j=1}^{k} y_j w_j \tag{C.4}$$

where k is the number of nearest neighbors, y_j is the target value of neighbor j, and w_j is a distance-related weight. The k parameter is typically learned during training.

Appendix C.4. Regression Tree

Building a decision tree can be thought of as recursively applying the process of dividing a single parent node into its two child nodes. As such, the process for the division of one node can be used to fully define the construction process. To find the data points that will be allocated between the two child nodes the optimal data partition is selected such that the sum of squares is minimized between the creation of the two new nodes, otherwise expressed as:

$$\arg\min \sum_{i=1}^{j} \sum_{k=1}^{N_i} (y_k - \bar{y}_i)^2 \tag{C.5}$$

where i is a new node, j is the total number of new new nodes, k is a data point in the i partition, N_i is the total number of data points in the partition i, and $\bar{y}_i$ is the average target value of the instances in partition i. A greedy algorithm is commonly used to select the partitions. The size of the regression tree (e.g. width and depth) is typically set prior to construction and a full tree is built in the first pass. Estimations are made by taking the average target value of all data points in a terminal node. Variance and complexity reduction is then achieved by pruning of the full tree. One possible regression tree cost function can be given as:

$$C_\alpha(T) = \sum_{m=1}^{T} \sum_{k=1}^{N_m} (y_k - \bar{y}_i)^2 + \alpha T \tag{C.6}$$

where α is a regularization parameter, N_m is the number of observations in terminal node m, and T is the number of terminal nodes. Minimization of equation C.6 is achieved through collapsing nodes to achieve a sub-tree T such that $T \subset T_o$ where T_o is the full tree. The weakest link approach to pruning is then commonly used: nodes are collapsed by order of least contribution to $C_\alpha(T)$ (i.e. lowest residual sum of squares error) such that a set of trees are produced that are gradually more generalized and can be fit to α.

Appendix C.5. Random Forest Regression

Random forest regression is an extension of regression trees that utilizes bootstrapping aggregation, that is, the aggregation of many "weak" regression trees into an ensemble model that is typically lower variance than its individual components. A random forest is constructed as:

1. For a training set of size n, set features X, and responses y, select β, the total number of trees in the forest.
2. Sample a subset of instances of the training data $X_t \subset X$, $y_t \subset y$ and train a new tree f_t on the data.
3. Repeat step 2 until β trees have been constructed.

To make an estimate using the random forest simply take the average prediction of all trees in the forest as:

$$\hat{f}(x^i) = \frac{1}{\beta} \sum_{i=1}^{\beta} f_t(x^i) \tag{C.7}$$

Appendix C.6. `XgBoost`

A complete mathematical description of `XgBoost` is beyond the scope of this work and a brief review is given here. `XgBoost` is an algorithm that applies the gradient tree boosting method. Gradient tree boosting takes advantage of both ensemble learners and iterative improvement to a model by means of using the residual loss of the previous iteration to train a new estimator. Estimators are added to the model such that the predicted result at iteration t is:

$$\hat{y}_i^t = \hat{y}_i^{t-1} + f_t(x_i) \tag{C.8}$$

where $f_t(x)$ is the new estimator. In the case of `XgBoost`, the generalized objective function to optimize during each step can be given as:

$$\text{obj}^{(t)} = \sum_{j=1}^{T} \left(G_j w_j + \frac{1}{2} H_j w_j^2 \right) + \Omega_{\lambda\gamma}(T) \tag{C.9}$$

where Ω is a model complexity contribution that is a function of the model of all the trees T with regularization parameters λ and γ. The parameter w_j is the leaf weights, and G_j and H_j are the sum of the first and second order components of the Taylor expansion of the specified loss function for leaf j, respectively. From here, an expression for a measure of how "good" a tree structure $q(x)$ is can be written as:

$$\text{obj}^* = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T \tag{C.10}$$

We can build a tree that continues to split nodes so long as the "goodness gained" from a given node split is larger than the regularization parameter γ, as:

$$\text{Gain} = \frac{1}{2}\left(\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right) - \gamma \tag{C.11}$$

where the components considered are the gain from the left leaf, the gain from the right leaf, and the score of the original leaf.

Appendix C.7. Gradient Boosting Regression

Gradient boosting regression relies on the construction of many weak prediction models that is built in an iterative process. The weak models in this case, and most often are, decision trees. A simplified description is provided here. First a base estimator, $F_0(x)$, is first trained on the data:

$$F_o(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma) \tag{C.12}$$

where L is a differentiable loss function. The pseudo-residuals, r_{im}, are then calculated for every data point i:

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right] \tag{C.13}$$

Another weak learner, $h_m(x)$, is then trained on the pseudo-residuals and the weight associated with the model, γ_m, is calculated using an optimization procedure. Not that here m indicates the current iteration of a total of M steps. The new model then has the form of:

$$\hat{F}(x) = \sum_{i=1}^{M} \gamma_i h_i(x) \tag{C.14}$$

New learners are then added to the model iteratively by re-calculating the pseudo-residuals to train a weak model and find it weights. This procedure can continue for an arbitrary number of steps or until a training metric passes a threshold.

Appendix C.8. Artificial Neural Networks

Fully dense artificial neural networks are composed of a network of many layers of interconnected nodes. Each node is connected to all nodes in the previous and following layers by a unique weight w. It is easiest to describe a neural network through the behavior of a single node - a diagram of which is provided in Figure C.18. The output of node i in layer j is calculated as:

$$a_{ij} = \sum_{i=1}^{n} w_i x_i + b \tag{C.15}$$

where b is a node-specific constant and there are n nodes feeding into the node a_{ij}. The value that feeds to all nodes in the following layer is then calculated as $f(a_{ij})$, where f is an activation function such as tan. For a regression problem there is often a final layer with a single node with no activation function to make predictions. The weights w and constants b are initially randomized and then learned through the training process. A mathematical description of the training procedure is beyond the scope of this work. The number of layers, activation function, and training procedure are all hyperparameters that can be optimized to best suit the problem at hand.

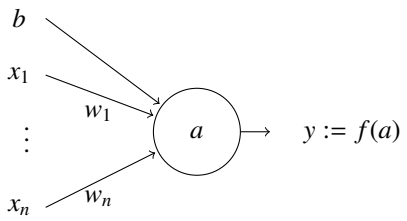


Figure C.18: Diagram of a single neuron in a neural network.

Appendix D. Performance of Machine Learning Models

The prediction accuracy of the various regression models can be shown by comparing the predicted output parameter versus the true input parameter for each simulation. This comparison is shown below in Figure D.19 for multiple regression models. Each subplot shows the author-selected input parameter data in red, and the randomly selected input parameter data in blue. Each subplot also has a dotted line with slope = 1 which can be used to determine the accuracy of each point. For a given simulation, if the predicted value matched the true value of the parameter, that point would fall on the dotted line. Therefore, any points in the upper triangle of a plot have an output parameter that is over-predicted by the regression model, and any points in the lower triangle have been under-predicted by the model.

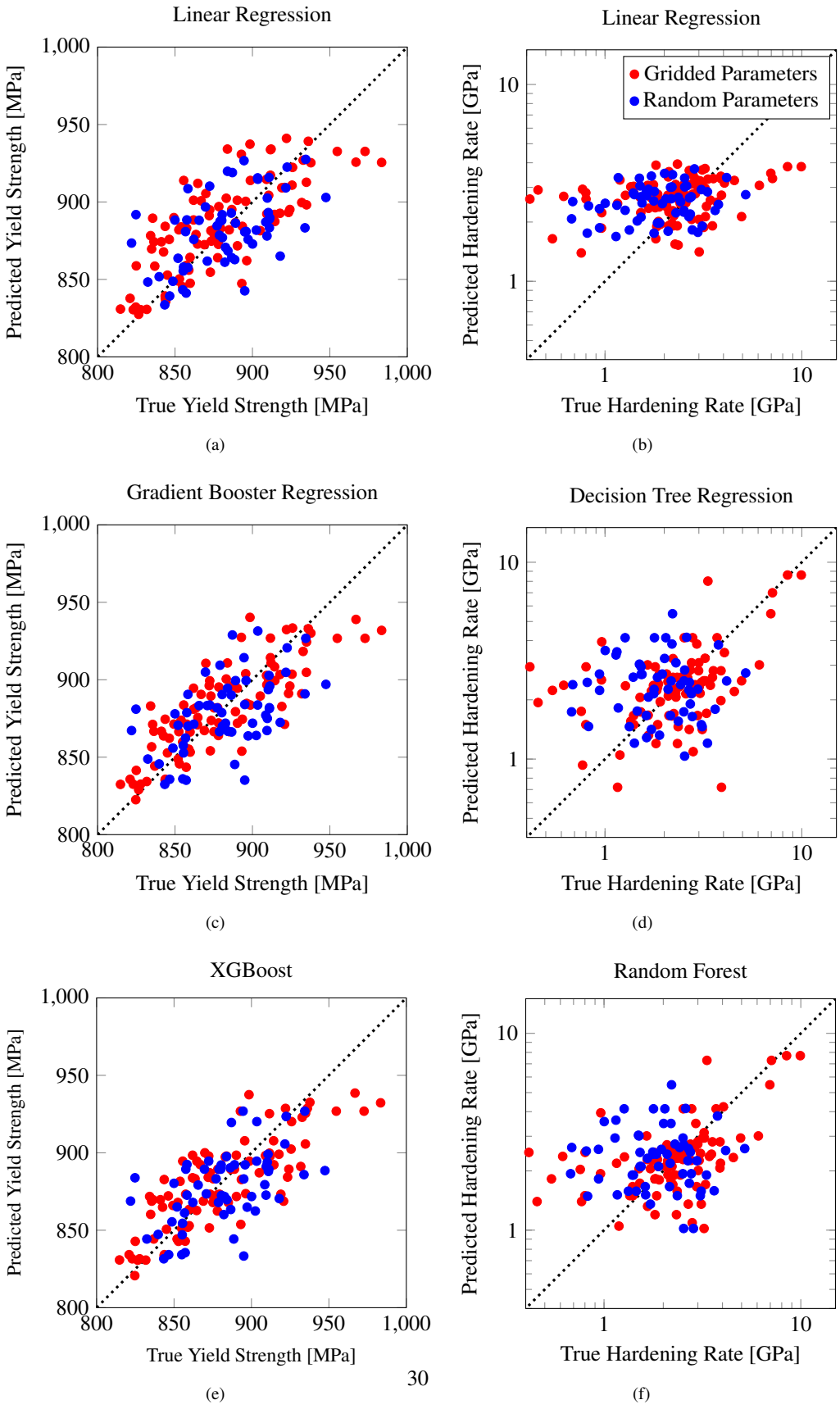


Figure D.19: True and predicted yield strength and hardening values for the the several models using the gridded and random test sets.