Pacific Northwest
NATIONAL LABORATORY

PNNL-37769

Materials Characterization, Prediction and Control Project

Summary Report on Data Analytics Framework

October 2024

K Kappagantula
M Obiri
MFN Taufique
L Richards
E King
S Howland
J Pope
L Durell
J Nguyen
J Koch
W Choi
JD Escobar Atehortua

With contributions from
DR Todd
H Kvinge
D Garcia
H Das
Y Guo
DM Brown
T Emerson
T Wang
KA Ross
M Pole
EI Barker
LE Smith

U.S. DEPARTMENT OF ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from
the Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062
www.osti.gov
ph: (865) 576-8401
fox: (865) 576-5728
email: reports@osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: http://www.ntis.gov

PNNL-37769

Materials Characterization, Prediction and Control Project

Summary Report on Data Analytics Framework

October 2024

K Kappagantula
J Pope
L Durell
L Richards
JD Escobar Atehortua
M Obiri
MFN Taufique
S Howland
J Nguyen

With contributions from
DR Todd
H Kvinge
D Garcia
H Das
Y Guo
DM Brown
T Emerson
T Wang
KA Ross
M Pole
EI Barker
LE Smith

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Abstract

This report summarizes the activities performed under the data analytics Vertex in the Materials Characterization, Prediction and Control Project funded under laboratory directed research and development at Pacific Northwest National Laboratory. The data analytics Vertex developed models for associating global or local process parameters, microstructural features, and performance properties of friction-stir-processed 316L stainless steel plates. Statistical, machine learning, and deep learning models, as well as generative artificial intelligence approaches, were used to develop the associations between the process-structure-property data streams. These associations formed the basis for predicting global properties of parts manufactured under different process envelopes, providing a basis for predicting performance using data-driven as well as physics-informed and physics-constrained approaches. Additionally, the associations were used to predict local process parameters and microstructural features of the product, predictive relationships that have the potential to form the basis of a control framework that could eventually modulate a friction-stir process to maintain product quality.

Executive Summary

The Pacific Northwest National Laboratory undertook the Materials Characterization, Prediction, and Control Laboratory Directed Research and Development Project to advance understanding of nuclear material processing and enable multifold acceleration in the development and qualification of new material systems produced via advanced manufacturing methods, such as solid phase processing, for use in national security and advanced energy applications (Smith 2021). This report provides details on two frameworks developed in the data analytics Vertex of the Materials Characterization, Prediction, and Control Project, titled PREDICT and CONTROL, which are composites of physics-constrained and data-driven models. The PREDICT framework associates global process conditions, microstructural features, and properties of friction-stir-processed (FSP) 316L stainless steel (SS) plates. The CONTROL framework predicts the local process parameters required to maintain microstructural features in a certain range that is desirable for maintaining part quality during manufacturing. Specific outcomes from this Vertex include:

1. A generative artificial intelligence model conditioned on microstructural images from the nugget region of the FSP 316L SS plates as a function of the global FSP parameters, namely tool traverse rate and tool temperature, which can classify conditioned images to relevant process conditions with 55—80% accuracy.
2. A microstructure segmentation tool based on the UNet++ architecture trained on correlated scanning electron microscope-electron backscatter diffraction images (developed exclusively for model development) consisting of labels for the grains and grain-boundary regions in the microstructural images of the nugget region of the 316L plates. The UNet++ architecture could achieve errors in predicting equivalent circular diameter as low as 0.5 μm in microstructures with average equivalent circular diameters of 1—5 μm.
3. Statistical, deep learning (DL) (symbolic regression) and machine learning (Random Forest, XGBoost) models to predict the average grain size of FSP SS grains as a function of global process parameters, showing a prediction accuracy of >80% (based on mean square error) even when trained on just 11 classes of experiments.
4. Statistical, DL (symbolic regression) and machine learning (Random Forest, XGBoost) models to predict the local average microhardness of FSP SS grains as a function of global process parameters showing a prediction accuracy of >96% (based on mean square error) even when trained on just 11 classes of experiments.
5. A physics-informed neural network, predicated on the universal differential equation paradigm, to predict the local tool temperature as a function of machine power, calculated as a function of the local torque and tool rotation rate by solving the lumped capacitance equation for a model dynamical system consisting of a FSP region.
6. A DL (long short-term memory) model to predict the tool rotation rate seen during FSP of 316L SS plates in the steady-state processing region as a function of the local process parameters experienced previously with an average error in prediction of tool rotation rate as low as 4 rpm over a range of 80—120 rpm.
7. A hybrid statistical-machine learning model associating local FSP parameters with mean average voltage obtained from non-destructively evaluating FSP 316L SS plates under steady-state conditions with a low mean average percentage error in prediction of ~4% that can capture the associations between the dataset within a 95% prediction interval.

Acknowledgments

This research was supported by the Materials Characterization, Prediction, and Control (MCPC) investment, under the Laboratory Directed Research and Development program at Pacific Northwest National Laboratory. Pacific Northwest National Laboratory is a multi-program national laboratory operated for the Department of Energy by Battelle Memorial Institute under Contract No. DE-AC05-76RL01830.

Authorship for this report reflects substantial contributions in developing models and/or technical leadership of MCPC data analytics Vertex activities throughout the duration of the project. Those responsible for specific characterization and experimental processing activities are identified in report attachments that describe that data. The authors wish to acknowledge the experimental team that performed tests that created the friction-stir processed material described in this report. The experimental team is listed in other MCPC Project literature.

Acronyms and Abbreviations

AC-WGAN	auxiliary classifiers-Wasserstein generative adversarial network
AI	artificial intelligence
BSE	backscatter electron
DA	data analytics
DL	deep learning
E2E	end-to-end
EBSD	electron backscatter diffraction
ECD	equivalent circle diameter
FSP	friction-stir processing
GAN	generative adversarial network
HAGB	high-angle grain boundaries
HD	Hausdorff distance
HV	Vickers hardness
ID	identification
LAGB	low-angle grain boundaries
LMM	linear mixed model
LSTM	long short-term memory
MAE	mean average error
MAV	mean absolute voltage
MCPC	Materials Characterization, Prediction, and Control
ML	machine learning
MLP	multilayer perceptron
MSE	mean square error
NDE	non-destructive evaluation
NVP	non-volume preserving transformations
OOB	out-of-bag
OOD	out of distribution
PCA	principal component analysis
PCC	Pearson Correlation Coefficient
PI	prediction interval
PINN	physics informed neural network
PNNL	Pacific Northwest National Laboratory
PSP	process-structure-property
Real-NVP	real-valued non-volume preserving transformations
RF	random forest

RFM	Random Forest Model
RGB	red, green, blue
RNN	recurrent neural network
rpm	rotations per minute
SE	secondary electron
SEM	scanning electron microscope
Seq2seq	sequence to sequence
ShAPE™	Shear Assisted Processing and Extrusion
SPP	solid-phase processing
SS	stainless steel
UT	ultrasonic testing
WGAN	Wasserstein generative adversarial network

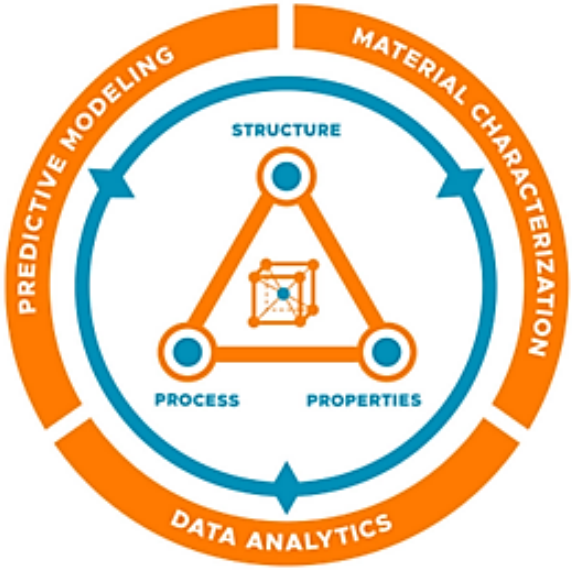
Contents

Figures

Tables

1.0 Introduction to Materials Characterization Prediction and Control Agile Investment

The Pacific Northwest National Laboratory (PNNL) undertook the Materials Characterization, Prediction, and Control (MCPC) Laboratory Directed Research and Development Project to advance understanding of nuclear material processing and enable multifold acceleration in the development and qualification of new material systems in national security and advanced energy applications (Smith 2021). The MCPC Project executed research across three scientific vertices—material characterization, predictive modeling, and data analytics (DA).



MCPC Project Logo

The central technical objective of the MCPC Project was to improve the prediction and characterization of the process-structure-property (PSP) relationships within the microstructurally refined region of stainless steel (SS) samples prepared using two solid phase processing (SPP) techniques. The two techniques, known as friction-stir processing (FSP) and Shear Assisted Processing and Extrusion (ShAPE™), are well-established at PNNL within the Solid Phase Processing capability through many years of investment across a range of materials and applications (PNNL 2024).

A lack of fundamental, science-based understanding of the material evolution hinders current efforts to adopt novel material processing approaches under various processing conditions. Without this fundamental understanding, the development and qualification of new material production systems rely too heavily on iterative, experimental approaches, resulting in long timelines that challenge product delivery. This stifles opportunities for the rapid uptake of high-performance advanced materials enabled by emerging fabrication. Through the MCPC effort, we aimed to transform how new material-processing technologies, such as the suite of SPP approaches, are understood and operationalized to enable multifold reductions in the timelines and costs for the development and qualification of new material production systems across several application areas. We developed a modeling and analysis framework that integrated accelerated predictive modeling, in-line monitoring of the process and product, and advanced DA. These frameworks will enable high-fidelity in-silico experiments, efficient design of experiments, and precise real-time process control which, taken together, will significantly accelerate development and qualification of materials systems when compared to traditional, iterative approaches.

2.0 Introduction to DA Vertex

Smart manufacturing integrates sensors and real-time monitoring to track product quality, process parameters, and energy usage, among other factors. DA plays a pivotal role in shaping the future of manufacturing by using data obtained during manufacturing from various sources, analyzing it, and providing data-driven insights. DA can enable advanced real-time monitoring of machinery, production lines, and supply chains. By analyzing data, we can identify bottlenecks, streamline workflows, and enhance the overall efficiency of a manufacturing process. Pyzer-Knapp et al. (2022) herald this approach to materials discovery and manufacturing as a part of the fourth paradigm of science, where “big data” along with the associated systems and analytics to process it are available. DA enables accelerating research and development cycles by establishing hitherto unknown relationships between process parameters, material microstructural features, and the properties of interest in the manufactured products. DA help maintain product quality by monitoring parameters such as temperature, pressure, and material consistency. It can detect defects early, reducing waste and ensuring consistent product quality.

Reliable data is essential for accurate analytics. Incomplete or inconsistent data can lead to flawed insights. Ensuring data quality and accessibility across different systems and sensors is a challenge in this environment, where smart manufacturing lines using DA are still an emerging epoch. Manufacturers often struggle with fragmented data across different systems. Integration and data harmonization are critical to unlock the full potential of analytics. Shared data ecosystems enhance collective insights. Successful implementation involves collaboration between data scientists, materials scientists, and production teams. Bridging knowledge gaps and fostering teamwork is essential.

The focus of the DA Vertex in the MCPC Agile is to analyze and interpret the PSP datasets developed on the FSP 316L SS plates to uncover patterns, trends, and correlations using data-driven approaches such as statistical analyses, machine learning (ML), and deep learning (DL) or artificial intelligence (AI) techniques. The relationships established between the process, structure, and property data streams are to form the basis for (1) accelerating research and development for discovering materials with desired properties or microstructural features or identifying properties and microstructural features of interest in specimens manufactured in a process envelope (2) consistently achieving desired part qualities. These focus areas are important, especially for manufacturing materials using advanced manufacturing approaches such as FSP and ShAPE™, where first-principles modeling approaches are either nascent or emerging but computationally heavy.

Discovering materials with desired properties (such as hardness, fatigue strength, creep strength, corrosion resistance, or ultimate tensile strength) or microstructural features (such as grain size, crystallographic orientation, degree of consolidation (% porosity), or precipitate density and morphology) requires the identification of processing envelopes, consisting of a set of parameters such as pressure, temperature, and feedstock composition that are required to synthesize the specimens. We note here that a framework that can predict process envelopes based on the desired properties or microstructures, can also determine the properties and microstructural features of a sample that is processed with a predefined set of global parameters. On the other hand, consistently achieving quality control requires the prediction of local process parameters, which result in parts or specimens with the desired properties or microstructures, as well as mechanisms for maintaining them during production/manufacturing. Activities in the DA Vertex were intended to either predict global processing conditions required to achieve a specimen with desired properties/microstructures, or control local processing

conditions to keep them in ranges known to achieve desired microstructures/performance. Specifically, for the MCPC Project, the predict module was developed to identify global process envelopes that can manufacture FSP 316L SS plates with high microhardness and low average grain size. The control module was developed to predict local tool temperatures in the transitory stage (i.e., plunge) and separately, tool rotation rates in the steady-state stage of FSP for 316L SS plates.

To establish correlations and make the predictions, we used statistical ML and DL models. The dataset used to develop these correlations was complex, comprising data as tables, graphs, images, and categorical labels, which were all accommodated in different modeling approaches. As such, the PREDICT and CONTROL modules are founded on frameworks that are composites of multiple models with significant data flow between them. The data flow for the PREDICT and CONTROL frameworks is provided in Figure 1. Only experimental data was used to assemble these frameworks in the current effort. However, the frameworks could easily extend to include computationally derived datasets in the future. The information flow shows the framework for predicting the microstructural features and properties of 316L SS plates manufactured at various global process conditions.

Dataset

The experimental dataset on which the PREDICT and CONTROL frameworks were developed consisted of FSP process parameters, microstructural images from scanning electron microscope (SEM) and electron backscatter diffraction (EBSD) modalities, an array of microhardness values in the cross-section of the FSP nugget region, and the mean absolute voltage (MAV) of the reflected ultrasonic signal in the shear wave mode (which is an indicator for the local grain size in the specimen). Table 1 provides a schematic representation of the various data streams used for developing the PREDICT and CONTROL frameworks.

Table 1. Various parameters used for PREDICT and CONTROL framework development using data-driven approaches in the DA Vertex.

Process Parameters	Structural Features	Properties
Traverse position Forge position Traverse velocity Forge velocity Traverse force Cross feed force Forge force Spindle speed Spindle torque Spindle power Traverse power Tool temperature	Grain size distribution Average grain size Grain boundary density Local voltage from ultrasonic transducer	Microhardness

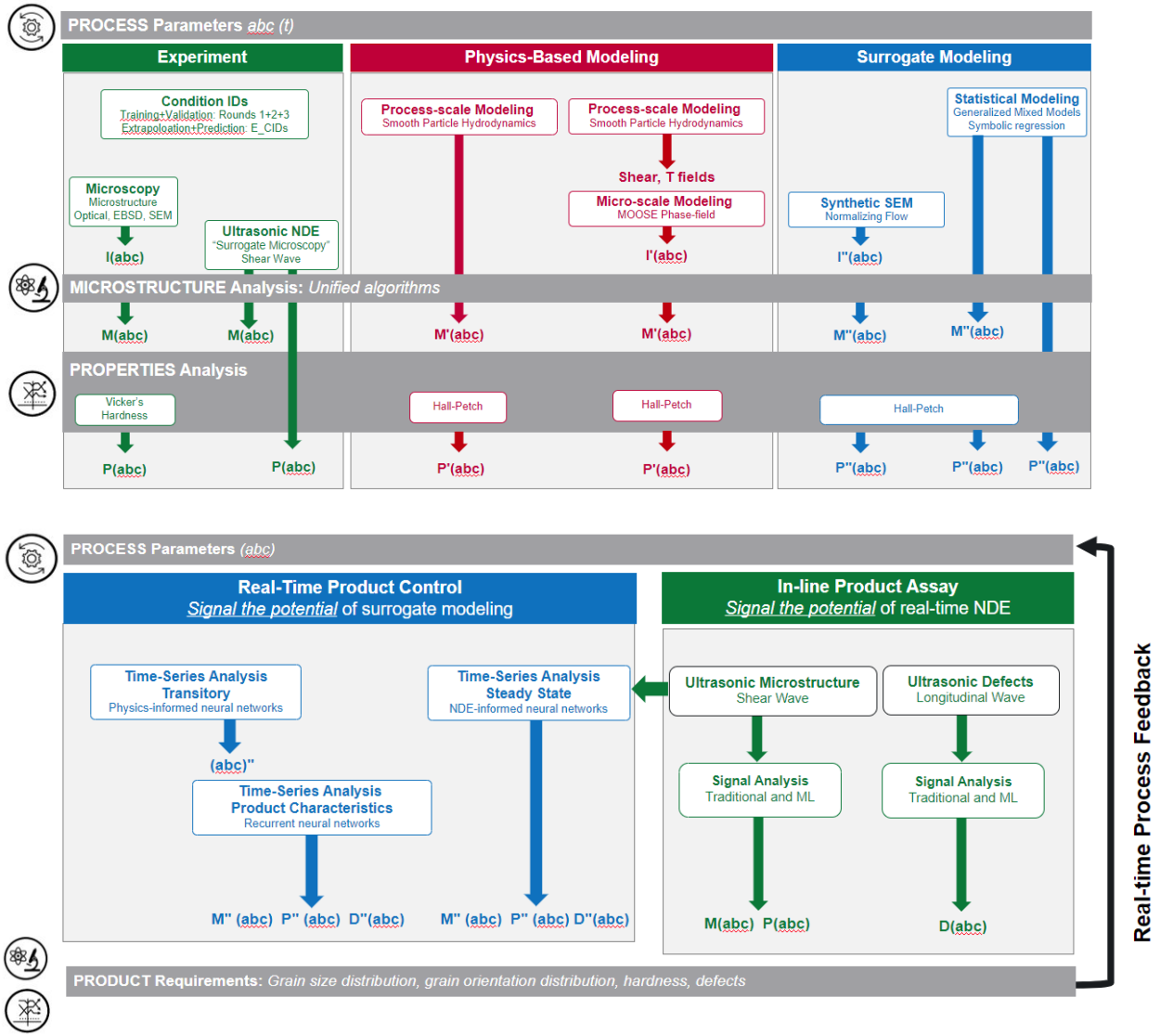


Figure 1. Information flow for the PREDICT (top) and CONTROL (bottom) frameworks with the DA Vertex activities represented in blue.

Table 2 provides the various global process conditions applied during FSP of the 316L plates. Of these, Conditions C01–C11 were performed initially, and the associated data was used for training the models. Subsequently, Conditions C12–C15 experiments were performed and the data from them was used to test the models.

Table 2. Experimental process parameters for all training and evaluation conditions.

Condition ID	Tool Temperature (C)	Traverse Velocity (in/min)	Traverse Velocity (mm/min)	Tool Force (lbF)	Tool Force (kN)
C00	N/A	N/A	N/A	N/A	N/A
C01	720	0.5	12.7	9,000	40.0
C02	720	0.5	12.7	10,500	46.7
C03	720	1	25.4	10,500	46.7
C04	720	1	25.4	9,000	40.0
C05	720	3	76.2	10,500	46.7
C06	750	1	25.4	10,500	46.7
C07	750	3	76.2	10,500	46.7
C08	800	1	25.4	10,500	46.7
C09	800	3	76.2	10,500	46.7
C10	850	1	25.4	10,500	46.7
C11	850	3	76.2	10,500	46.7
C21*	**800**	**4**	**101.6**	**10,566**	**47.0**
C22*	**850**	**2**	**50.8**	**10,566**	**47.0**
C23*	**900**	**3**	**76.2**	**10,566**	**47.0**
C24*	**900**	**2**	**50.8**	**10,566**	**47.0**

* New experiments called end-to-end (E2E) run after C01-C11 were completed.

Each local process parameter combination is also accompanied by corresponding local process conditions collected through the sensors attached to the FSP machine during sample synthesis. Details of these local process parameters, as well as the sensors used to collect them, are available in Appendix 1. Additional information on the microstructural features and properties obtained through destructional testing, as well as the ultrasonic test data obtained through non-destructive evaluation (NDE) of the samples, is available in multiple reports corresponding to the round of characterization (Todd et al. 2024a, 2024b, 2024c, 2024d).

2.1 Objectives

The focus of the DA Vertex is to develop the PREDICT and CONTROL frameworks for material discovery and process control during advanced manufacturing processes. To achieve these outcomes, we identified the following objectives:

1. PREDICT: Develop a framework for associating process-structure-properties of FSP 316L plates.

    a. *Predict FSP 316L microstructural features as a function of global process parameters used to manufacture the samples.*

        i. Image based predictions:

1) Generate conditioned microstructural images of FSP 316L samples as a function of the global process parameters used to manufacture them.

2) Segment the generated microstructural images to determine the grain size distribution in the microstructures of FSP 316L plates.

ii. Numerical predictions:

1) Predict average grain size of FSP 316L steel plates as a function of global process parameters applied to manufacture them.

b. Predict the microhardness properties of FSP 316L plates as a function of the global process parameters used to manufacture them.

2. CONTROL: Develop a framework of associating material properties or microstructural features or their surrogates with local process conditions.

a. Predict local tool temperatures during FSP of 316L plates as a function of the spindle power under unsteady processing conditions.

b. Predict local tool rotation rate at a point in time during FSP of 316L plates as a function of the local process parameters applied previously under steady-state conditions.

c. Predict average grain size of processed specimens in the steady-state conditions as a function of process parameters and ultrasonic testing voltage.

Each of these objectives were achieved in a specific task, as listed below in the DA Vertex.

Objective #	Objective	Associated Task
1.a.i.1)	Generate conditioned microstructural images of FSP 316L samples as a function of the global process parameters used to manufacture them.	Generative AI for Developing Synthetic Microstructural Images Conditioned on Process Parameters
1.a.i.2)	Segment the generated microstructural images to determine the grain-size distribution in the microstructures of FSP 316L plates.	Unified Microstructural Analysis - Deep Learning Models for Grain Size
1.a.ii.1)	Predict average grain size of FSP 316L steel plates as a function of global process parameters applied to manufacture them.	Process-Structure Relationship Modeling
1.b	Predict the microhardness properties of FSP 316L plates as a function of the global process parameters used to manufacture them.	Process-Property Relationship Modeling
2.a	Predict local tool temperatures during FSP of 316L plates as a function of the spindle power under unsteady processing conditions.	Physics Informed Neural Networks for Transitory Parameter Prediction
2.b	Predict local tool rotation rate at a point of time during FSP of 316L plates as a function of the local process parameters applied previously under steady-state conditions.	Recurrent Neural Networks for Steady State Parameter Prediction

Objective #	Objective	Associated Task
2.c	Predict average grain size of processed specimens in the steady-state conditions as a function of process parameters and ultrasonic testing voltage.	Grain Size Prediction using Ultrasonic Testing Data

2.2 Scope

This scope provides the approaches adopted for developing the models for the PREDICT and CONTROL frameworks of the DA Vertex. Specifically, we provide the method for model development, along with the results and a preliminary discussion of the findings—including model uncertainty quantification in certain cases. Section 3.0 discusses the development of the PREDICT framework, while Section 4.0 outlines the procedures and results from the CONTROL framework. In Section 5.0, we discuss the conclusions from the present efforts and provide a view on the future work in this area.

3.0 Developing Framework for MCPC Characterize and Predict

In this section, we present the methods for developing the PREDICT framework based on which process parameters, microstructural features, and properties may be predicted given one of the other data streams for FSP 316L steel plates. We describe the outcomes from three tasks focusing on generating synthetic microstructural images corresponding to specific FSP global process conditions, segmenting microstructural images (derived from experimental practices or generative models), and developing process-structure and process-property associations between the numerical global FSP process conditions and the mean/average grain size (signifying the FSP 316L plate microstructural feature of interest) or local/average microhardness of the processed plates. Results from the framework development, namely the model associations demonstrated by the accuracy of the predictions in the subsequent sections, specifically address Objective A. The central idea of the PREDICT framework is that it is viable to predict the global process parameters, microstructural features, or properties of FSP 316L plates with sufficient information that can aid in several subsequent activities. Principally, our interest is to develop the central associations that will be the backbone for statistical ML or DL models that can support the design of experiments in advanced manufacturing assignments. The first two tasks address the challenge of using image and numerical data in combination for predicting microstructural features when process condition data is specified. The third task addresses the challenge of predicting average microstructural features or average properties of a manufactured specimen, when the global parameters used to manufacture it are known.

3.1 Generative AI for Developing Synthetic Microstructural Images Conditioned on Process Parameters

There is a need for anticipating the microstructures of metallic specimens manufactured by a specific approach under predetermined conditions. Such images provide an idea of the features that can be anticipated in the microstructures, such as grain size distribution, composition, texture variation, among others. Understanding the spatial distribution of such features provides powerful information on the plausible performance of the specimen. In some cases, such capability can preclude process envelopes that may result in features and their distributions that are not desirable owing to certain restrictions on properties or performance.

Generative adversarial networks (GANs) are generative models trained using a competitive “game” between two modules (Goodfellow et al. 2014). A generator model learns to transform random noise samples into datapoints in a target data distribution of interest, while a discriminator model learns to differentiate between ground truth datapoints and synthetic ones produced by the generator. This adversarial game is decoupled from any specific generator or discriminator architecture, allowing for the use of highly expressive models. GANs are known for producing high-fidelity images in many domains, though their hard-to-sample and, at times, ill-formed latent spaces (Liu and Regier 2020) can make it difficult to achieve the levels of consistency and verifiability desirable for many scientific applications.

Wasserstein GANs (WGANs) introduced a new formulation based on minimizing the Wasserstein-1 distance in order to remedy the original GANs’ unstable training dynamics. This approach used weight clipping to enforce 1-Lipschitz continuity on its discriminator module, which is a key constraint and assumption of the WGAN loss function. This weight clipping turned out to limit the underlying models’ expressiveness and was later replaced with a gradient penalty on the loss to preserve both high performance and stable training. Normalizing flows are

invertible neural networks known for their utility in generative and density estimation tasks. They perform generation by transforming samples from a simple base distribution—commonly a multivariate standard distribution—into the more complex data distribution of interest. This is similar to how GANs function, though normalizing flows possess unique advantages on account of their stipulated invertibility and efficient computation of their Jacobian determinants. These constraints allow normalizing flows to directly perform maximum likelihood optimization rather than the adversarial game employed by GANs. While normalizing flows have lower sample quality than GANs on many traditional red, green, blue (RGB) image domains, their improved training stability, sample efficiency, and recall, which are crucial elements for scientific applications where data is expensive to acquire.

There are several normalizing flow layers in use to address the joint constraints of invertibility and tractable Jacobian determinants. Many prominent approaches use coupling layers based on non-linear independent component estimation, real-valued non-volume preserving transformations (Real-NVP), and neural spline flows. We leverage Real-NVP for our models in this work due to the technique's widespread availability and enduring use in normalizing flow research.

3.1.1 Methods

In this subtask, we leveraged two classes of SEM images collected from FSP 316L steel plates. Both classes of SEM images were collected from samples subjected to 11 different FSP conditions and having a different average Vickers hardness value. The first class of SEM images focused on quantity over quality, gathering multiple SEM scans from a wide variety of locations in the cross-section of the FSP samples. As such, the first class of images produced a larger amount of visual data, though the individual steel grains are not crisp. The second class of SEM images focused on quality over quantity, gathering crisper images from a smaller number of sample locations, also in the cross-section of the sample.

In order to balance capturing multiple grains in a single sample and having data samples that can be tractably leveraged using common graphical processing unit hardware, we created chips from all the SEM images from first and second classes, using a sliding window of size 256 × 256 pixels and an overlap of 80% between consecutive windows. We derived >2-million chips (2,486,400 chips to be specific) from the first class of SEM images and 74,646 from the second class of images. The greater number of Class 1 chips—33 times those from Class 2—enable us to evaluate how data quantity and quality affect different families of generative models.

We tracked the process parameters, namely forge force, tool traverse velocity, and steady-state temperature observed during the FSP experiments for each chip. With our Class 2 data, we also tracked the magnification level used to image the FSP 316L microstructure. These features make up the conditioning data we used to guide each model's sampling, allowing us to target combinations of process parameters and magnification levels at inference time. For evaluation, we also tracked the average hardness of the sample associated with each chip. We list all the experimental conditions used in our training and evaluation in Table 2, and hereafter refer to combinations of experimental process parameters by their associated experimental condition identification (ID) (e.g., Condition 3).

We compared two families of generative models, GANs and normalizing flows, on their ability to capture both the visual characteristics of our steel SEM images and their underlying process parameters and hardness. Specifically, we leveraged WGANs with gradient penalty loss and ResNet-18 auxiliary classifiers (AC-WGAN) to incorporate conditioning information, as well as

Real-NVP. All models were implemented using Pytorch, with evaluation performed using Lightning.

When training our models, our goal was to compare the GAN and normalizing flow approaches at their best, given the same compute budget and training data. All models were trained for four days on a single Nvidia V100 graphical processing unit. Under the V100's memory constraints, we could train our GANs with a batch size of 64 while our normalizing flows could only manage a batch size of 24. Both GANs and normalizing flows work by transforming a multivariate Gaussian noise sample into a synthetic datapoint, but the increased memory consumption (and, by extension, smaller batch size) for our normalizing flow stems from the fact that GANs can work with noise samples of arbitrary dimensionality, whereas normalizing flows must use noise samples with the same dimensionality as the training data to preserve their invertibility. Specifically, we found a 100-dimensional standard normal to be effective for our GANs, while our normalizing flows were constrained to a noise distribution of dimension 65,536 (256 × 256) to match the size of our SEM chips. Within the four-day training time, our GANs could train on roughly 64 million datapoints, while the normalizing flows trained on just over 5 million.

To evaluate our synthetic microstructures, we sampled 400 random microstructures from experimental data and those generated by AC-WGAN and Real-NVP models, proceeding to capture the mean and standard deviations of the first four moments of their chips' pixel intensity distributions. Specifically, these moments are the mean, standard deviation, skew (asymmetry of pixel intensity), and kurtosis (degree of outlier pixel intensities). These moments allow us to better assess the properties of each data distribution and their global alignment with one another with no full manual inspection or biases from human visual intuition. Because of space limitations, for these evaluations we report results for only a single experimental process parameter configuration: tool temperature of 720 °C, tool traverse velocity of 1.0 in/min, forge force of 10,500 lbF, and average sample hardness of 266.12 HV. The results and analysis for this condition are represent our analysis of all 11 experimental conditions.

We also finetuned ResNet-18 visual classifiers pretrained on ImageNet to assess how well our generative models capture process parameter and sample hardness information in their synthetic microstructures. We finetuned one ResNet-18 each on 90% of datasets for a single epoch, then evaluated the remaining 10% of the corresponding experimental data as well as 400 synthetic samples from each experimental condition drawn from both AC-WGAN and Real-NVP. Because of the limited number of discrete values each process parameter takes, we cast predicting them as a classification rather than a regression problem. Because the spread of average sample hardness values is much more diverse, we instead trained the ResNet-18 models to minimize the L1 regression loss. We arrived at finetuning ResNet-18 for a single epoch after comparing it to finetuning and from scratch training of ResNet-18, ResNet-50, and ResNet-152 models for 1, 2, 4, and 8 epochs. For space, we reported only this configuration as it proved the most effective approach.

3.1.2 Results and Discussion

Figure 2 shows five microstructure chips from Class 1 and 2 SEM images, respectively. We also show the synthetic microstructural image chips generated by the AC-WGAN and Real-NVP models. For both SEM classes, we see that AC-WGAN does not meaningfully capture either the grain structures or the diversity of overall shades in the experimentally obtained SEM images. In particular, the grain distributions produced by AC-WGAN were spotty and disconnected against a noisy background. Real-NVP-generated images were more diverse in terms of both intra-chip grain structure and inter-chip intensities, though there is still room to improve the crispness of

the grain boundaries. From a visual perspective, Real-NVP presents closed loop grains in the synthetic images, albeit more predominantly when trained on Class 2 SEM images. We performed hyperparameter optimization over several axes following the initial experiments. Specifically, we looked at learning rate magnitude and scheduling, gradual introduction of the auxiliary classifier loss to help improve the network's training dynamics, as well as differing rates of generator/discriminator training and network sizes to account for underfitting or overfitting. We found these results from the reported configuration to be the most competitive, with visual performance degrading as model size increased *or* decreased. Our ability to experiment as broadly with Real-NVP's size was limited because of its memory consumption, though visual results improved as we increased the model capacity to the reported configuration, which hit the upper limits of our compute capability.

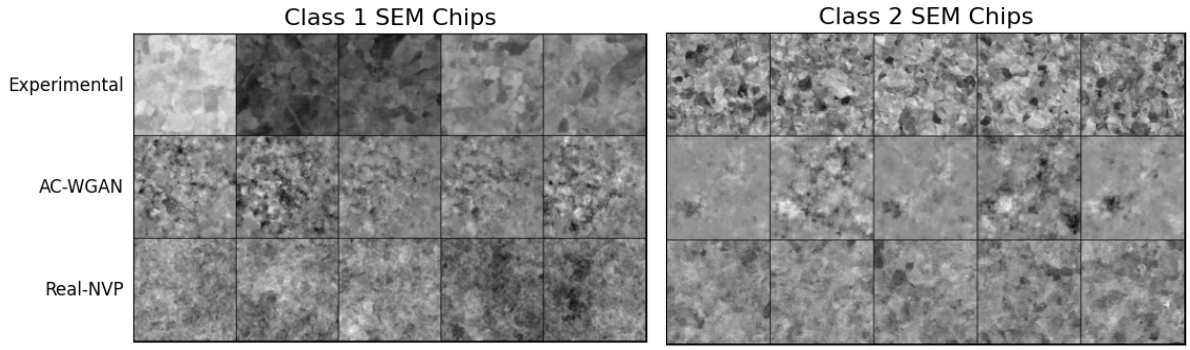


Figure 2. Schematic showing representative images obtained experimentally, and those generated by AC-WGAN and Real-NVP models trained exclusively on Class 1 and 2 SEM images separately, corresponding to the same process Condition C10.

We collected image statistics for a broader sampling of chips across all models and data conditions to better characterize model performance at the more complete *distributional* level. Figure 3a and b visualize the distribution of pixel counts across all sampled experimental and generative chips. The Wasserstein distance between the experimental and AC-WGAN image pixel intensity was 77.57, while that between experimental and Real-NVP generated images was 35.27 for the images trained on Class 1 SEM images. For the models trained on Class 2 images, the Wasserstein distance between the image pixel intensity of experimental and AC-WGAN was 100.33, while that between experimental and Real-NVP images was 64.92. Table 3 shows statistics over 500 randomly sampled image chips from our process parameter condition. Specifically, we capture the first (mean), second (standard deviation), third (skew or asymmetry), and fourth (kurtosis or outlier) moments of the chips' pixel intensity distributions. A more negative skew shows images are darker than they are bright and vice versa, while a higher kurtosis value shows a higher prevalence of unusually dark or bright pixels relative to most of the image. Both skew and kurtosis are unitless measures, and so we compare relative to experimental statistics to interpret them. Pixel intensities are integers ranging from [0, 255].

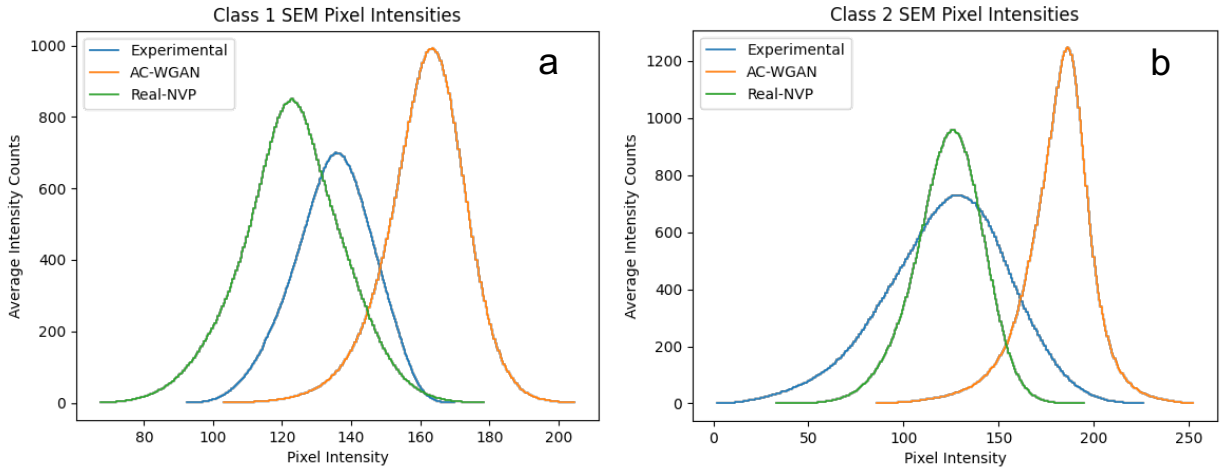


Figure 3. The average pixel intensity counts for experimental data and generative models trained on Class 1 and 2 experimental SEM images.

The statistics corresponding to Class 1 SEM images paint a picture partially at odds with our initial visual inspection. There is only one out of the four moments—the standard deviation of pixel intensities—where Real-NVP meaningfully outperforms AC-WGAN in similarity to experimental data. In particular, Real-NVP's substantially lower mean pixel intensity shows that it could be disproportionately sampling the darker chips seen in Figure 2; further sampling and manual inspection revealed this to be the case. These differences are captured in Figure 3, which shows AC-WGAN's pixel intensities being well-centered with experimental data, with Real-NVPs being a better fit overall aside from their erroneous mean. By contrast, the images from models trained on higher-quality Class 2 data saw far stronger results from Real-NVP, which was significantly closer to experimental statistics across all four modes, having tighter standard deviations than AC-WGAN in all cases. Figure 3 also shows Real-NVP's pixel intensities as far better aligned with experimental data than those of AC-WGAN, though both methods still fail to capture the full diversity and statistical spread of experimental chips.

Table 3. Image statistics for experimental, AC-WGAN, and Real-NVP microstructure chips for models trained on Class 1 and 2 SEM images along with the mean and standard deviation of the first (mean), second (standard deviation), third (skew), and fourth (kurtosis) moments of each chip's pixel intensity distribution for comparison. The model whose statistics most closely align with experimental data for each moment are bolded.

Class 1 Model	Mean	Standard Deviation	Skew	Kurtosis
Experimental	165.20 ± 24.0	13.75 ± 1.20	-0.31 ± 0.23	3.20 ± 1.30
AC-WGAN	**160.43 ± 2.87**	15.06 ± 3.67	**-0.44 ± 0.13**	**4.10 ± 0.48**
Real-NVP	126.38 ± 4.87	**12.97 ± 1.47**	-0.1 ± 0.20	**3.70 ± 2.21**

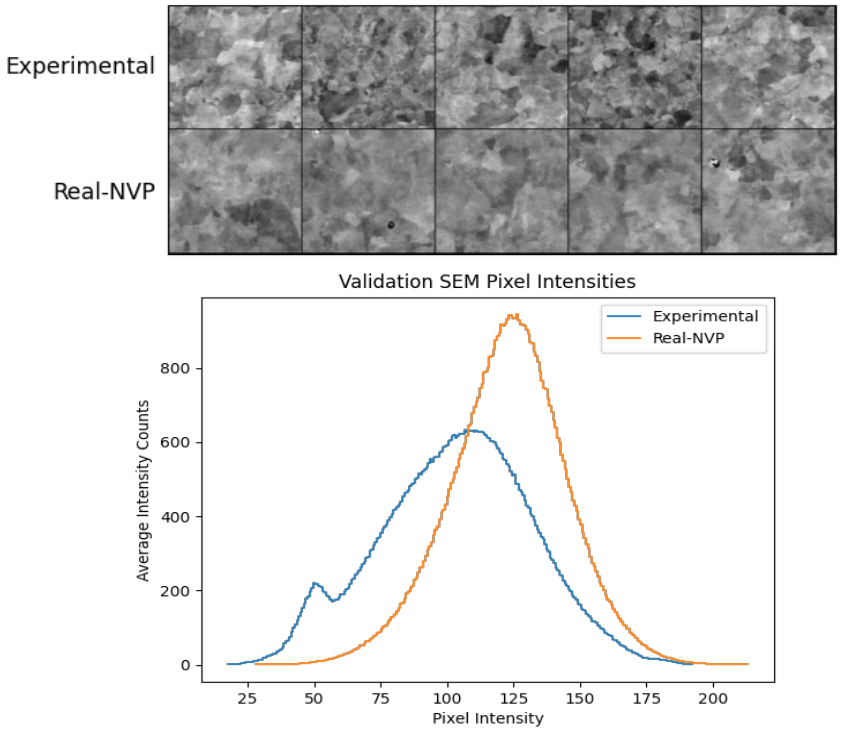
Class 2 Model	Mean	Standard Deviation	Skew	Kurtosis
Experimental	133.71 ± 3.16	30.63 ± 1.48	-0.31 ± 0.12	3.05 ± 0.22
AC-WGAN	160.09 ± 7.99	16.37 ± 4.66	-0.62 ± 0.33	6.46 ± 2.09
Real-NVP	**127.26 ± 2.52**	**18.24 ± 1.06**	**-0.41 ± 0.16**	**3.63 ± 0.79**

Table 4 shows the ResNet-18 classification and regression scores for process parameters and material properties on Class 1 and 2 SEM images data across our experimental and generative sources. To account for class imbalance in the process parameters, we included a trivial baseline that simply predicts the most common label for reference. Results that outperform this trivial baseline are bolded for clarity. We see that even making the concession of treating our process parameters as discrete classes rather than continuous labels, estimating process parameters and sample hardness from SEM chips alone is challenging—even when operating over experimental data. Indeed, our finetuned ResNet struggled to perform above a trivial baseline on most conditions. This is likely because of a combination of class imbalance among our experimental conditions and low effective data counts overall; while we have a fair number of chips for both rounds of data, we had to use an 80% chip overlap to reach those numbers means that the number of wholly unique chips is actually several factors lower than the final data counts. We note here that decreased overlap likely would not improve performance, especially on prediction (as opposed to generation). This is both because overlap discourages overfitting for models and because decreasing overlap does not improve the underlying image quality. This has implications for the use of process parameters and material property estimation as a tool for evaluating models in advanced manufacturing spaces.

Table 4. ResNet-18 classification and regression scores for process parameters and material properties on Round 2 and 3 data across our experimental and generative sources.

		Accuracy of Experimental Global Process Parameter Prediction		
Dataset	**Chip Source**	**Forge Force**	**Traverse Velocity**	**Tool Temp.**
	Trivial Classifier	81.8%	45.5%	45.5%
Class 1 images	Experimental	**90.0%**	40.0%	40.0%
	AC-WGAN	**83.3%**	**50.0%**	33.3%
	Real-NVP	81.8%	18.1%	**45.5%**
Class 2 images	Experimental	**83.4%**	**50.0%**	32.6%
	AC-WGAN	**83.3%**	**50.0%**	22.4%
	Real-NVP	78.6%	44.3%	**55.0%**

Given that the visual inspection and image statistics revealed that the Real-NVP generated synthetic SEM images of the FSP 316L plate microstructures were more representative of the experimental images compared to those generated by AC-WGAN, we chose to further develop the Real-NVP model in this task. This choice was also supported by the accuracy of the ResNet-18 classification model for the Real-NVP images based on global tool temperatures used during processing, which has a significant effect on the grain size distribution seen in SEM images. We generated synthetic SEM chips corresponding to the end-to-end (E2E) Condition C12 and compared them visually with the experimental images and associated statistics from FSP 316L steel plates subjected to Condition C12 in order to test the efficacy of the Real-NVP model. Figure 4 shows the comparison of images and schematics of Real-NVP generated images corresponding to Condition C12, and the associated experimental data. The Wasserstein distance between synthetic and experimental pixel intensities was 85.53. We note that there is a close agreement between the visual images and the image statistics as well.



End-to-End	Mean	Standard Deviation	Skew	Kurtosis
Experimental	109.36 ± 10.8	27.16 ± 4.0	-0.02 ± 0.44	3.28 ± 1.28
Real-NVP	133.48 ± 3.8	18.57 ± 2.5	-0.24 ± 0.22	3.82 ± 1.52

Figure 4. Schematic showing representative images obtained experimentally, and those generated by Real-NVP model trained on E2E experimental images, corresponding to the same process Condition C12 (top); the average intensity counts of images from E2E experimental SEM chips and Real-NVP generated synthetic chips as a function of pixel intensity (middle); and a table with image statistics for experimental and Real-NVP microstructure chips for models trained on E2E images along with the first (mean), second (standard deviation), third (skew), and fourth (kurtosis) moments of each chip's pixel intensity distribution for comparison.

Our results of this task show that normalizing flows provide an appealing alternative to GANs for characterizing FSP 316L steels. We observe more sophisticated and, particularly with models trained on higher-quality Class 2 SEM data, much more statistically realistic visual microstructure generation using normalizing flows. We suspect that the main reasons for this stem from our problem space. Specifically, the small scale of manufacturing data, coupled with the difficulty of transfer learning between large-scale, RGB object classification datasets to microstructure imagery. In these circumstances, normalizing flows' superior sample efficiency and training stability are able to shine. We also suspect that normalizing flows' incorporation of conditioning information into their coupling layers and not requiring an additional discriminator loss term as with GANs, avoids incentives in optimization that encourage GANs to produce visually incoherent outputs that contain information that can differentiate between different experimental conditions. However, the same difficulty in transferring large, pretrained RGB vision models to our datasets also inhibits our ability to evaluate how well our models capture conditioning and downstream material property information. Adapting vision models pretrained on ImageNet proved disappointing, and our own datasets were too small to effectively train such models from scratch. While visual and statistical characterization of synthetic data is a good

start and provides useful insight, in the future it will be important to apply better methods for direct evaluation of synthetic microstructures' material relevance, consistent with metrics commonly used for experimental images (e.g., grain-size distribution).

3.2 Unified Microstructural Analysis - DL Models for Grain Size

The microstructure of a material plays a crucial role in determining its physical, mechanical, and functional properties. In FSP, the processing parameters employed during manufacturing largely influenced grain size (Garcia et al. 2024). Notably, localized variations in microstructure can occur for a static set of processing parameters (Wang et al. 2024). The variation might be because of the presence of a temperature and strain gradient through the material because of the tooling geometry (Liu and Nelson 2017). The unprocessed material in front of the tool undergoes rapid deformation, leading to grain fragmentation via the generation of low-angle grain boundaries (LAGBs) and the deterioration of pre-existing annealing twin boundaries. However, within the stirring microstructure, the combination of strain, strain rate, and temperature leads to dynamic recrystallization, reducing LAGBs and promoting twinning (Liu and Nelson 2016, 2017).

The final microstructure of FSP austenitic SSs is complex, and therefore, EBSD is needed to achieve a complete understanding of the crystallographic evolution induced by the process. For example, the coexistence of low and high angle grain boundaries, coincidence site lattice Σ3 annealing twin boundaries, and dense dislocation walls, results in a rich variety of observable features when using conventional electron imaging modes, such as secondary electrons (SE) and backscatter electrons (BSE). The inherent microstructural variability within FSP microstructures requires large-area EBSD characterization. However, accurate EBSD data is closely related to time consuming measuring parameters, such as diffraction pattern collection time, frame averaging, and step size (i.e., pixels per map), resulting in long data collection times, especially when large areas of analysis need to be characterized. With SPP austenitic SS, a fine step size EBSD map (i.e., 50-100 nm) takes longer to capture than a comparable SE/BSE image over the same area owing to the differences in the data collection approach. This happens as the microstructure is composed of very fine grains containing intra-granular stored strain, which reduces the quality of the diffraction patterns. Acquiring high-quality data on such microstructures is complex since for every pixel in the image, an EBSD pattern needs to be measured during a specific collection time, then averaged over several repetitions, and finally computationally indexed against a crystallographic database before moving to the next position. EBSD data quality is strongly susceptible to surface quality after standard metallographic preparation, representing a labor-intensive and expertise-dependent task. Efforts in high-throughput experimentation and quality assurance/quality control have motivated exploration into advanced image analysis techniques, focusing on DL, to streamline the process of microstructure characterization.

DL offers advanced toolsets to enhance and automate the analysis of microscopy images (Ede 2021). Semantic segmentation, in particular, has showed a use for high-throughput materials characterization (Ede 2021, Morgan et al. 2022). Though standard segmentation model architectures are typically applied to microscopy data, the collection of large volumes of labeled microscopy images on which to train a model is often an issue. In our previous work, multiple segmentation architectures were employed on SEM images of 347H SSs manufactured by a melting-cold rolling-annealing based approach, with labels provided by EBSD, to segment grain boundaries and quantify grain size distributions (Chowdhury et al. 2024). Annealing of the samples produced a clear visual distinction between the grains and grain boundaries, leading to impressive model performance. However, FSP produces a more complicated microstructure,

where the austenitic grains are substantially finer and contain considerable amounts of LAGBs and dense dislocation walls.

Internal strain produces high intragranular contrast in SEM images, which impedes quantitative microstructural analysis. Generating labeled images for training is itself a challenging task. To address impediments to strained image analysis, we examined the coupling of different modalities (BSE and SE) to EBSD measurements in order to establish high-quality labeled data on which to train semantic segmentation models to identify grain boundaries in the current effort. We then used the ensemble of models trained on the best performing modality to segment a series of BSE images of samples manufactured with different FSP parameters.

3.2.1 Methods

This task leveraged SE and BSE images collected using a JEOL IT500 HRLV field emission SEM. A region of interest was fixed, and images were collected at 10 and 20 kiloelectron volt (keV) accelerating voltages, both in SE and BSE modes. EBSD was collected using a JEOL 7600 SEM, equipped with an Oxford Synergy CMOS EBSD detector. An acceleration voltage of 20 keV and a step size of 500 nm were used in all cases. A pixel-to-pixel correlation between SE, BSE, and EBSD was possible by maintaining a constant region of interest using fiducial marks visible in all imaging modes.

DA was performed using MTEX 5.8.2 toolbox in Matlab R2020b (Bachmann et al. 2011). A five-pixel neighbor clean was used to denoise random electronic noise and local mis-indexing zones. Then, a misorientation threshold of $\omega = 15°$ was used to identify high-angle grain boundaries (HAGBs), following the conventions used by Jazaeri and Humphreys and Humphreys (Jazaeri and Humphreys 2004; Humphreys 2001). Whereas lLAGBs were identified for a misorientation window of $2° < \omega \leq 15°$. The remaining intra-granular boundary information was classified as dense dislocation walls, i.e., $\omega \leq 2°$, and was excluded from the LAGB and HAGB reconstruction sequence. Finally, annealing twin boundaries, i.e., coincidence site lattice $\Sigma 3$, were reconstructed using the methodology proposed by Patala et al. (2012). After boundary reconstruction and identification, LAGBs, HAGBs, and twins were deconvoluted, labeled, and then extracted to generate independent boundary-specific skeleton-like plots, as shown in Figure 5.

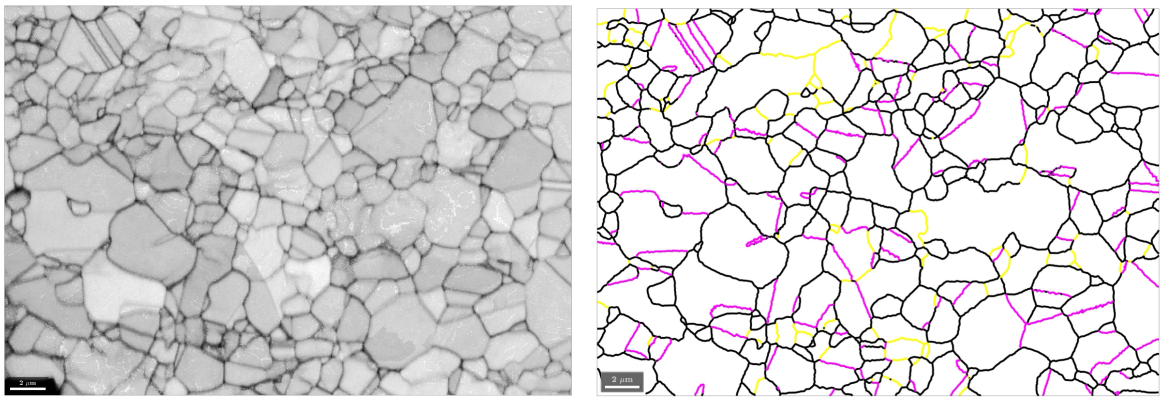


Figure 5. EBSD band contrast image and the corresponding boundary-specific skeleton plot for a FSP 316L plate microstructure processed at 850 °C, 1.0 in/min traverse rate, and a forge force of 10,500 pound (corresponding to the C10 condition).

Prior to microstructural characterization, platinum fiducials were deposited on the stir zone of the four out of distribution (OOD) samples at five different locations per sample, as shown below in Figure 6. BSE images were captured at each location using a Helios Hydra UX dual-beam plasma focused ion beam/SEM at an accelerating voltage of 20 keV, working distance of 4.0 mm, 0° tilt, and 5,000× magnification. The BSE images were cropped to 1024 × 1024 pixels for prediction by our segmentation models. EBSD data were acquired using a plasma focused ion beam, equipped with an Oxford Synergy CMOS EBSD detector. EBSD data collection was carried out at an acceleration voltage of 20 keV, with a step size of 100 nm consistently applied across all measurements using AZtec software. AztecCrystal was used to generate inverse pole figure maps, image quality maps, and evaluate grain sizes. The mean equivalent circle diameter (ECD) was calculated for each sample location using a threshold angle of 5° and minimum grain diameter of 0.4 µm.

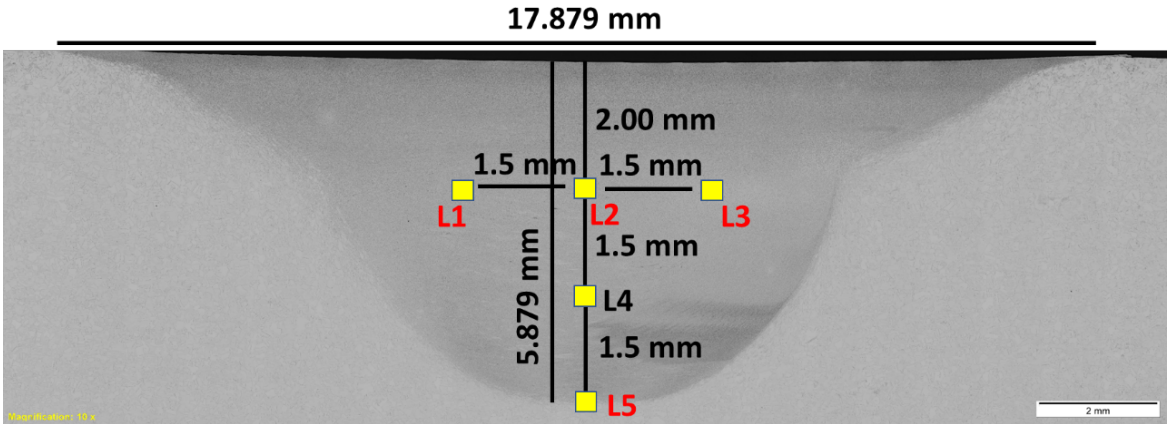


Figure 6. Locations at which BSE and EBSD measurements were taken for the OOD samples.

Semantic segmentation requires labeled training data, meaning that the input image must have a corresponding segmentation map with a pixel-to-pixel match. Hand-drawn segmentation maps are often necessary to create labeled training data for microscopy images. Because FSP causes the formation of many dislocations in the 316L austenitic SS microstructure, the high amount of intragranular contrast in their SEM images complicates the accurate manual identification of grain boundaries. Therefore, we performed sequential, correlated SEM and EBSD measurements to produce labeled training data.

To create the training data, the ground truth grain boundary map (output) was obtained from the EBSD boundary reconstruction. LAGB, HAGB, and twins were grouped into a single grain boundary class. Factors such as the stage tilt complicated the registration of the SEM images and grain boundary maps and drift distortion that obscure or distort features in the EBSD map, requiring specialized post-processing procedures. Differences in tilt angles between SEM and EBSD lead to foreshortening of grains and varying interaction volumes, while differences in magnification and working distance led to varying image resolution, and differences in accelerating voltage and beam current lead to varying probe sizes—all of which require correction to obtain accurate pixel-to-pixel matches (Tong and Britton 2021). We first registered the fiducial marks in the EBSD grain boundary map and SEM images, then manually adjusted the EBSD grain boundary map to obtain pixel-to-pixel correspondence. This process resulted in four sets of training pairs, as shown in Figure 7.

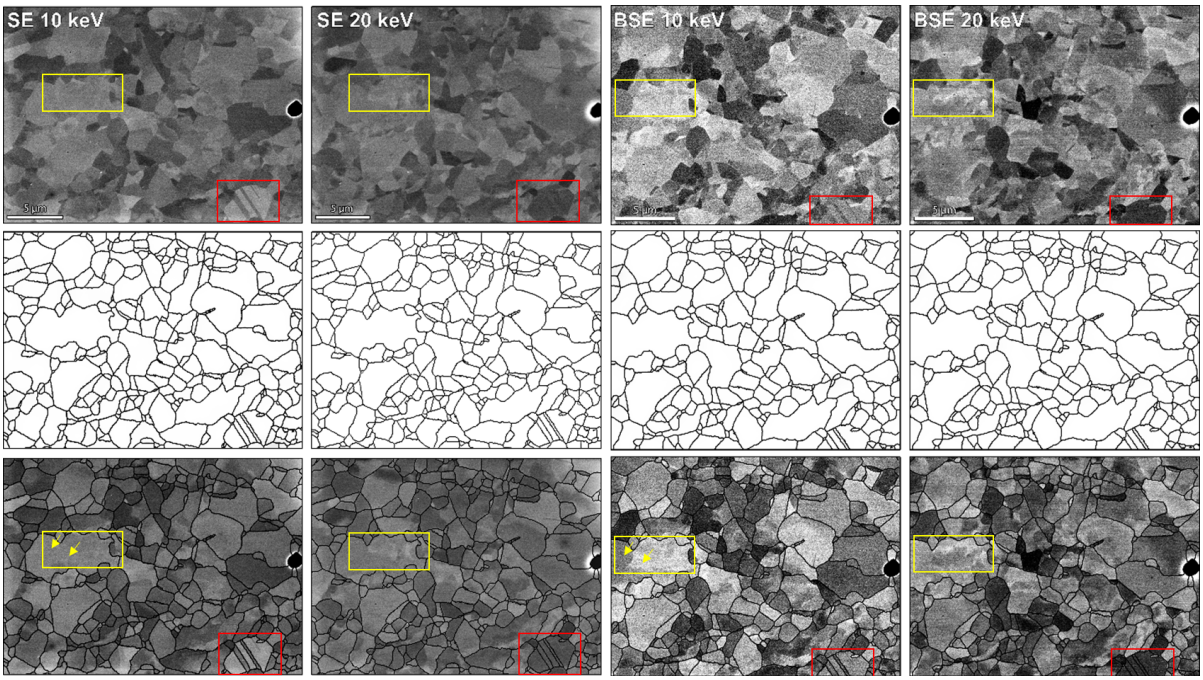


Figure 7. SEM images, registered EBSD-generated grain boundary maps, and overlays after registration.

We used the UNet++ architecture (Zhou et al. 2018) pretrained on either the common ImageNet dataset or MicroNet, a large dataset of over 100,000 labeled microscopy images (Stuckner et al. 2022), to train semantic segmentation models to identify grain boundaries in SEM images. Identifying grain boundaries in an SEM image is a highly class-imbalanced problem. The grain boundary class accounts for <10% of all pixels in an SEM image, while the rest is made of the grain class. A topological loss function, TopoLoss (Hu et al. 2019), was added to the DICE loss function to train the models. The dual loss function is shown in Equation (1).

$$Loss_{TopoDICE} = Loss_{DICE} + \lambda Loss_{Topo} \quad (1)$$

where $Loss_{DICE}$ is the same DICE loss, $Loss_{Topo}$ is the topological loss, and λ is a hyperparameter that controls the relative weighting between the two losses. An Adam optimizer, an algorithm for first-order gradient-based optimization of stochastic objective functions, based on adaptive estimates of lower-order moments, with an adaptive learning rate from $1e^{-4}$ to $1e^{-5}$ was used to train the dual loss function.

The images used in this subtask were 1792 × 1280 pixels. Each registered SEM–EBSD pair was split into three smaller rectangle patches of 256 × 1280 pixels to create three distinct test sets: test set 1 contains chips A1 to A5, test set 2 contains chips D1 to D5, and test set 3 contains chips E1 to E5. This allows us to (1) perform threefold cross-validation for each image modality and (2) generate an ensemble of models trained on each modality. For each set, the remainder of the image was split into non-overlapping square chips of 256 × 256 pixels to create the training set, shown in Figure 8. For predicting the grain boundary map of test set 1, columns B to G were used as training data, while column A was used as test data. For predicting the grain boundary map of test set 2, columns A to C and E to G were used as training data, while column D was used as test data. For the grain boundary map of test set 3,

columns A to E and G were used as training data, while Column F was used as test data. During training, standard augmentation methods were employed by applying horizontal flips, vertical flips, and 90° rotations to the chips. Each augmented training set was then divided into 90%/10% train/validation splits. A total number of 4 SEM–EBSD pairs were collected for this subtask, resulting in 464 chips available for model development. A total of 105 train and 11 validation sample chips, along with the aforementioned rectangular test sample, were obtained for each set, resulting in a total number of 116 chips for training and a rectangular strip equivalent in size to 5 chips for testing each model.

We used the F_1 score to measure the performance of supervised semantic segmentation models. The F_1 score is the harmonic mean of precision and recall and is computed as follows in Equation (2):

$$F_1 = \frac{2TP}{2TP + FP + FN}, \tag{2}$$

where TP is the number of true positives (correctly classified pixels), FP is the number of false positives (pixels incorrectly classified as belonging to the grain boundary), and FN is the number of false negatives (pixels incorrectly classified as not belonging to the grain boundary). Note that grain boundaries are present on the atomic scale and are not directly observed at the resolutions of images in this work. When we refer to a pixel belonging to a grain boundary, this shows that the area contained within that pixel is adjacent to and/or overlapping with the physical grain boundary. The thickness (pixel width) of what we call the grain boundary is, therefore, arbitrary. False positives in pixel-wise segmentation because of thicker boundaries are not as meaningful as false positives away from grain boundary regions.

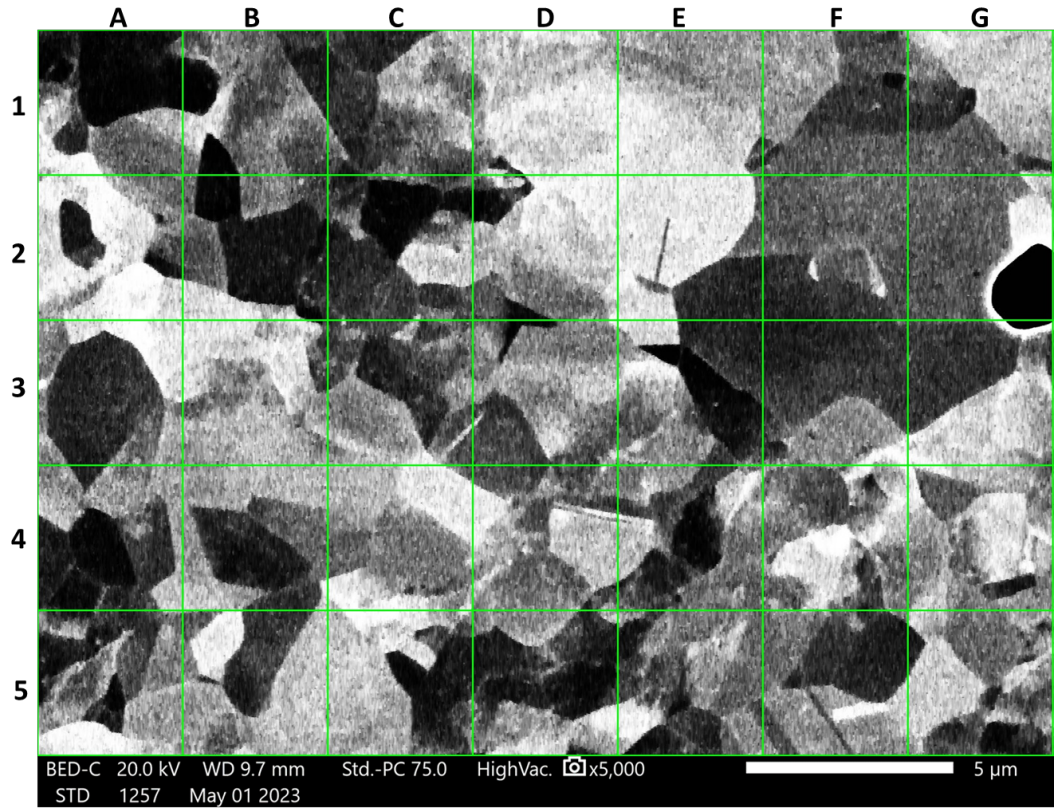


Figure 8. Location of the training set splits. The green lines show individual chips, while the labels show the columns used for the corresponding test sets.

In addition to the F_1 score as a general measure of model performance, it is useful to quantify the extent of the segmentation error. The Hausdorff distance provides such a measure by capturing the maximum discrepancy between two corresponding images (i.e., the segmentation map and the grain boundary map). In our application, the Hausdorff distance represents the maximum distance of a false positive prediction to the nearest grain boundary pixel. Because the Hausdorff distance is highly sensitive to outliers, we apply the 95% percentile Hausdorff distance (HD95), which excludes the top 5% of outliers (Jia et al. 2022).

Equiaxed grains are often described by the ECD, which is the diameter of a circle with the same area as the grain. We automated grain identification and measurement from a grain boundary map using scikit-image and the micron-to-pixel conversion factor (C). For the grain boundary map produced by EBSD, C is derived from the EBSD measurement. For the grain boundary maps aligned with the BSE and SE micrographs, C is derived from the micrograph.

3.2.2 Results and Discussion

Table 5, Table 6, and Table 7 show the average performance metrics of the three models, namely the ImageNet with DICE, MicroNet with DICE, and MicroNet with TopoDice for each modality and model pretraining/loss scheme. The F_1 scores, HD95, and mean average error (MAE) in the mean ECD for individual models trained on each modality are presented below as well.

Table 5. Average threefold crossvalidation performance of models trained on different image modalities and with pretraining/loss schemes.

Image Modality	F_1 Score			HD95			MAE in ECD (μm)		
	ImageNet DICE	MicroNet DICE	MicroNet TopoDICE	ImageNet DICE	MicroNet DICE	MicroNet TopoDICE	ImageNet DICE	MicroNet DICE	MicroNet TopoDICE
BSE 10	0.56	0.58	0.57	49.4	44.5	48.3	1.49	1.42	1.75
BSE 20	0.50	0.51	0.50	55.7	55.2	53.7	1.32	1.10	1.23
SE 10	**0.60**	**0.62**	**0.62**	**30.6**	**26.3**	**28.7**	**1.08**	**0.68**	**0.57**
SE 20	0.52	0.52	0.53	35.9	36.1	34.9	1.86	1.15	1.29

Table 6. Performance metrics for each UNet++(ImageNet/DICE) model finetuned on the specified training image and set.

Training Image	Training Set	F1 Score	HD95	Number of Grains			ECD (μm)		
				Pred	True	% Error	Pred	True	Error
BSE 10 keV	set1	0.61	30.1	3	17	82%	3.37	1.34	2.03
	set2	0.51	77.9	6	11	45%	2.85	1.91	0.94
	set3	0.55	40.3	4	12	67%	3.25	1.76	1.49
BSE 20 keV	set1	0.54	36.4	6	17	65%	3.19	1.34	1.85
	set2	0.44	77.4	6	11	45%	2.92	1.91	1.01
	set3	0.51	53.1	7	12	42%	2.85	1.76	1.10
SE 10 keV	set1	**0.65**	24.3	17	28	**39%**	2.35	1.84	**0.51**
	set2	0.61	**19.1**	11	30	63%	3.21	1.77	1.44
	set3	0.53	48.3	10	22	55%	3.43	2.14	1.29
SE 20 keV	set1	0.49	29.5	6	27	78%	3.95	1.86	2.09
	set2	0.57	22.2	8	31	74%	3.79	1.75	2.04
	set3	0.49	55.9	8	23	65%	3.57	2.12	1.45

Table 7. Performance metrics for each UNet++(MicroNet/DICE) model finetuned on the specified training image and set.

Training Image	Training Set	F1 Score	HD95	Number of Grains			ECD (μm)		
				Pred	True	% Error	Pred	True	Error
BSE 10 keV	set1	0.65	28.5	7	17	59%	2.70	1.34	1.36
	set2	0.53	69.9	5	11	55%	3.13	1.91	1.22
	set3	0.57	35.1	4	12	67%	3.43	1.76	1.67
BSE 20 keV	set1	0.57	35.0	10	17	41%	2.08	1.34	0.74
	set2	0.45	84.7	4	11	64%	3.74	1.91	1.83
	set3	0.52	45.8	8	12	33%	2.48	1.76	0.72
SE 10 keV	set1	**0.67**	**16.0**	19	28	**32%**	2.39	1.84	**0.55**
	set2	0.62	20.1	18	30	40%	2.38	1.77	0.61
	set3	0.56	42.9	13	22	41%	3.01	2.14	0.87

Training Image	Training Set	F1 Score	HD95	Number of Grains			ECD (μm)		
				Pred	True	% Error	Pred	True	Error
SE 20 keV	set1	0.48	35.8	11	27	59%	2.99	1.86	1.13
	set2	0.61	19.0	13	31	58%	2.97	1.75	1.22
	set3	0.48	53.7	9	23	61%	3.22	2.12	1.11

Comparing across image modalities, we see models trained on the SE images taken at an accelerating voltage of 10 keV (SE 10) performed best across all models over all three metrics. Across each modality, models pretrained on MicroNet outperformed those pretrained on ImageNet. The benefit of the addition of topoloss to the loss function (denoted TopoDICE) is unclear, with performance improving and worsening across different metrics and different training images. TopoDICE enhances accuracy in ECD when model performance is better but decreases accuracy if a certain performance threshold cannot be met with DICE alone. The topological loss rewards conformity in the number of continuous, enclosed areas in the ground truth and predicted grain boundary maps, without considering the actual location of the grain boundary pixels. Conversely, DICE rewards pixel-level overlap and does not consider continuity. Therefore, we hypothesize that if pixel-level overlap cannot be accurately learned, rewarding continuity only further decreases accuracy.

For the best performing modality (SE 10), TopoDICE reduced the MAE in ECD from 0.68 μm to 0.57 μm but gave the same the average F_1 score of 0.62, though the average HD95 slightly increased from 26.3 to 28.7. Because our target task is to characterize grain structure, we weight MAE in ECD higher than HD95 and consider our best set of models to be the MicroNet/TopoDICE model trained on the SE 10 keV image.

To understand why model performance was highest when trained on the SE 10 keV image, we must examine the physical mechanisms behind image collection modes and acceleration voltages. Elastic scattering produced BSE images between the beam and the sample, while SE images are the product of inelastic scattering of electrons originated much closer to the surface relative to BSE. In a solid solution microstructure, such as for an austenitic SS, orientation contrast and electron channeling effects caused image contrast differences within the microstructure in BSE mode (Goldstein et al. 2017, Lloyd 1987). Therefore, for our 316L SS samples, contrast variations in the BSE images (produced at a fixed beam acceleration voltage) are mainly a reflection of the various grain and subgrain orientations, as well as small misorientation regions separated by dense dislocation walls, originating during dynamic recrystallization via SPP. SE images are much more sensitive to surface topography, and therefore subgrains and dense dislocation walls are less pronounced, resulting in reduced intragranular contrast and less noise (see Figure 7). It is important to notice that SE images are still somewhat sensitive to crystallographic contrast, and therefore become useful for training tasks aimed at grain boundary detection while minimizing the contribution of features produced by residual strain.

It should be clarified that the appearance of both SE and BSE images will be altered if the acceleration voltage is changed (i.e., interaction volume). This implies that there is not a unique SE or BSE image that defines a ground truth, but a compendium of multiple images that can fundamentally represent the same microstructure. Consequently, the generation of a ground truth image (SE or BSE) must be labeled with the acceleration voltage at which it was generated. In the case of backscattered electrons, our Monte Carlo simulations performed for an equivalent 316L SS solid solution and a beam normal to the surface, show escape depths

(interaction volumes) of approximately 70 nm and 250 nm, for beam energies of 10 keV and 20 keV, respectively. During EBSD, however, the sample is tilted to 70° relative to the horizon, further reducing the interaction volume to values within 50-100 nm (Nowell et al. 2005). Based on these concepts, it becomes intuitive that an accurate electron image that pairs an EBSD map should be collected at a reduced acceleration voltage to match a similar interaction volume. This is evident for BSE images but should also be considered for SE images, which are still mildly sensitive to crystallographic contrast. The red rectangles in Figure 7 are useful to show the effect of interaction volume on a series of annealing twins inside one austenitic grain. By reducing the acceleration voltage to 10 keV during imaging (relative to 20 keV for EBSD), these twin boundaries become consistent with the EBSD observations, thus providing a better match between electron image and crystallographic ground truth.

In our training sample, the mean ECD determined by EBSD is 2.55 μm, but individual grains with ECDs <1 μm were observed, and dislocation walls were present throughout grains at even finer length scales. Therefore, at each pixel, incident beam electrons have the chance to encounter a variety of submicron structures not related to HAGBs or LAGBs. For example, regions of the microstructure highlighted by yellow rectangles in Figure 7 show intragranular features associated mainly with stored strain and dense dislocation walls. Such regions may reflect in false positive grain boundary identifications (see yellow arrows in Figure 7), especially in BSE images. This effect can be exacerbated by a larger interaction volume since the resultant image is an average of features coming from different depths (see red squares in Figure 7).

In summary, the reduced sensitivity of SE images to crystallographic contrast and electron channeling contrast, alongside the shallower interaction volumes involved during image generation (relative to BSE), result in a more balanced image. Since SE images are also less sensitive to stored strain but are simultaneously sufficiently sensitive to grain and subgrain contrast, these become more useful for pixel-to-pixel correspondence with EBSD crystallographic ground truth with less propensity for false positive boundary identification.

Because our models were trained on images from a sample produced under a single set of FSP conditions, namely Condition C10, we then investigated the models' ability to generalize two samples produced under different FSP conditions. Segmentation models for microscopy images are known to generalize poorly to OOD images because of a variety of imaging and material parameters (Hirabayashi et al. 2024). To assess the OOD performance of our models, we examined a set of 20 BSE images with differences from the training set in terms of both material processing conditions and imaging parameters (different instrument, different instrument operator, different modality) as noted in Section 3.2. Table 8 provides the comparison of imaging parameters across the images collected for model training and OOD validation. The only commonality among the OOD images and all training images is the material (316L SS). The BSE training sets share the same imaging modality as the OOD set, and the BSE 20 keV training set also shares a common accelerating voltage.

Table 8. Comparison of imaging parameters across training and OOD images.

Image Set	Processing Condition	Imaging Modality	Accelerating Voltage (keV)	Resolution (μm/px)
Training	C10	BSE	10	0.010
Training	C10	BSE	20	0.010
Training	C10	SE	10	0.015

Image Set	Processing Condition	Imaging Modality	Accelerating Voltage (keV)	Resolution (μm/px)
Training	C10	SE	20	0.015
Training	C10	EBSD	20	0.016
OOD	C21, C22, C23, C24	BSE	20	0.027

Figure 9 shows a representative BSE image from the OOD set overlaid with segmentation maps for the MicroNet/TopoDICE model trained on the SE 10 keV image, along with the corresponding grain detections. The individual models produced segmentation maps with gaps in grain boundaries, which leads to erroneous grain detection and artificially increases the measured grain diameters. Ensembling the predictions by summing segmentation maps from the three models trained on the same modality led to improved grain boundary closure and, thus, more accurate grain detection. Figure 9 demonstrates the poor grain boundary closure for individual sets and the improvement gained with ensembling.

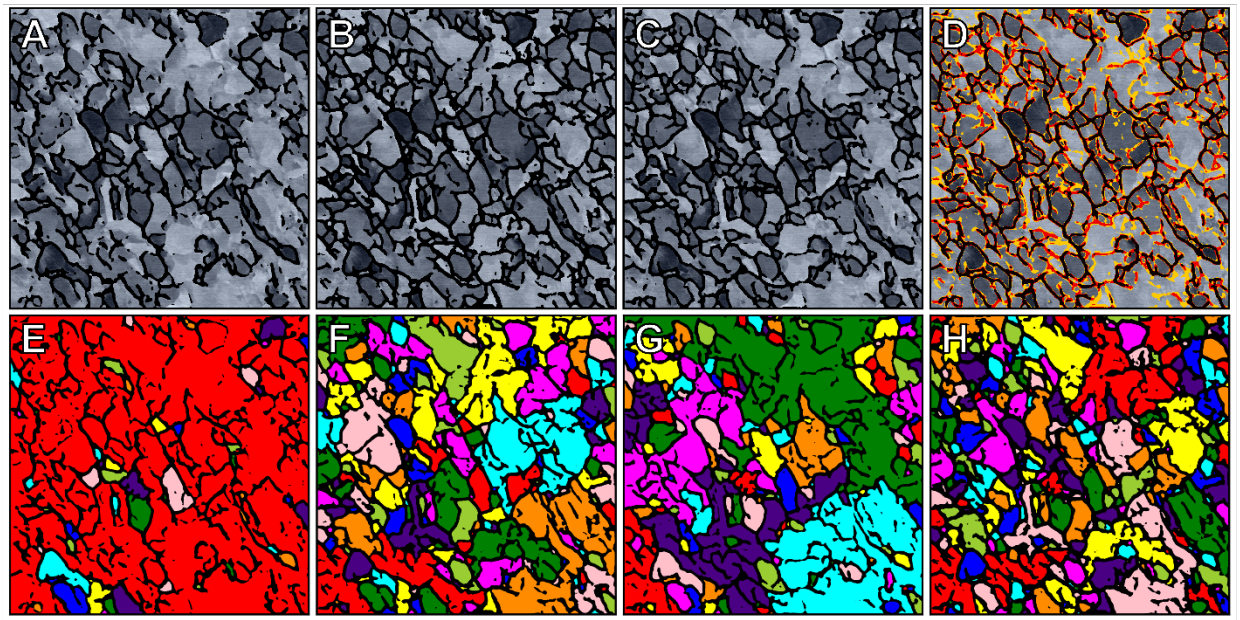


Figure 9. (A-D) BSE images and overlaid segmentation maps produced by the UNet++/MicroNet/TopoDICE/SE_10 model for an OOD sample and (E-H) corresponding grain detections for (A,E) training set 1, (B,F) training set 2, (C,G) training set 3, and (D,H) ensembled output of all training sets.

Table 9 and Table 10 show the MAE (μm) in the ECD from segmentation maps and the mean number of grains for each modality training set and ensemble. In each case, ensembling recovers more grains, which improves the accuracy of grain diameter measurements. Going forward, we apply ensembling for each model training modality when predicting segmentation maps for the OOD images. Because we do not have pixel-to-pixel alignment between BSE images and EBSD measurements of the OOD samples, we validate model performance through comparison of mean ECD determined from the ensembled segmentation maps and EBSD measurements. It should be noted that a perfect match in ECD between the two modalities is not expected because of differences in interaction volume, as discussed above, as well as the measurement technique. For instance, BSE will not distinguish grains with the same crystallographic structure and different orientations, while EBSD will. Moreover, because LAGBs

can be filtered from grain boundary maps produced by EBSD, grain diameters in materials with strong crystallographic texture can be larger when determined by EBSD compared with BSE (Humphreys 2001). With these differences in mind, we expect to observe similar statistical trends in mean ECD rather than one-to-one sample correspondence (Humphreys 2001).

Table 9. MAE (μm) in the ECD from segmentation maps predicted from the OOD image set against the grain diameter measured by EBSD in the corresponding region.

		Training Set Modality			
Model	Training Set	BSE 10 keV	BSE 20 keV	SE 10 keV	SE 20 keV
ImageNet/DICE	set 1	4.74	4.66	3.11	4.80
	set 2	3.25	3.29	2.86	4.81
	set 3	3.59	4.05	2.47	5.28
	ensemble	2.04	2.49	0.87	0.85
MicroNet/DICE	set 1	6.88	5.23	3.13	5.35
	set 2	8.05	6.15	0.76	5.37
	set 3	7.42	7.04	0.54	6.78
	ensemble	19.40	11.64	0.40	2.07
MicroNet/TopoDICE	set 1	24.41	14.09	7.93	6.97
	set 2	27.09	18.26	0.85	20.35
	set 3	25.84	20.25	1.84	12.38
	ensemble	21.00	9.59	0.34	4.37

Table 10. Mean number of grains from segmentation maps predicted from the OOD image set against the grain diameter measured by EBSD in the corresponding region.

		Training Set Modality			
Model	Training Set	BSE 10 keV	BSE 20 keV	SE 10 keV	SE 20 keV
ImageNet/DICE	set 1	4.65	4.50	18.95	14.35
	set 2	8.65	8.70	21.50	13.75
	set 3	7.55	5.45	21.80	12.40
	ensemble	48.75	42.70	90.70	87.80
MicroNet/DICE	set 1	1.85	4.05	17.70	12.00
	set 2	1.05	2.95	47.30	11.05
	set 3	1.25	1.85	38.20	12.25
	Ensemble	6.05	14.70	123.90	74.20
MicroNet/TopoDICE	set 1	3.00	11.80	27.75	40.15
	set 2	1.20	7.95	88.25	7.85
	set 3	2.10	5.40	71.25	26.20
	ensemble	5.30	20.25	127.45	60.20

Table 11 gives the overall MAE in ECDs obtained from the ensembled segmentation maps over the full OOD sample set. The MicroNet/TopoDICE model trained on the SE 10 keV image gives the lowest MAE of 0.34 μm, followed by the MicroNet/DICE model trained on SE 10 keV of 0.40 μm. The MAEs for the MicroNet models trained on BSE 10 keV and 20 keV images were extremely high because of the drastic underprediction of the grain boundary pixels.

Table 11. MAE in the ECD (μm) from the ensembled segmentation maps predicted for OOD image set against that measured by EBSD in the corresponding region.

	Training Set Modality			
Model	**BSE 10 keV**	**BSE 20 keV**	**SE 10 keV**	**SE 20 keV**
ImageNet/DICE	2.04	2.49	0.87	0.85
MicroNet/DICE	19.40	11.64	0.40	2.07
MicroNet/TopoDICE	21.00	9.59	**0.34**	4.37

Based on the MAE, it appears the ImageNet/DICE model provides more consistent, albeit less accurate, predictions across training image modalities. However, further examination reveals that the ImageNet/DICE model produces a narrow range of ECD values across the OOD samples. Figure 10 shows the individual predictions on OOD samples across models trained on the SE 10 keV images. Table 12 provides ECD distributions for the remaining training set modalities. EBSD-determined ECDs range from 1.47–4.68 μm, while ImageNet/DICE/SE 10 keV ECDs range from 3.07–4.40 μm. The MicroNet models more closely reproduce the expected ECD distribution, especially in the lower size range.

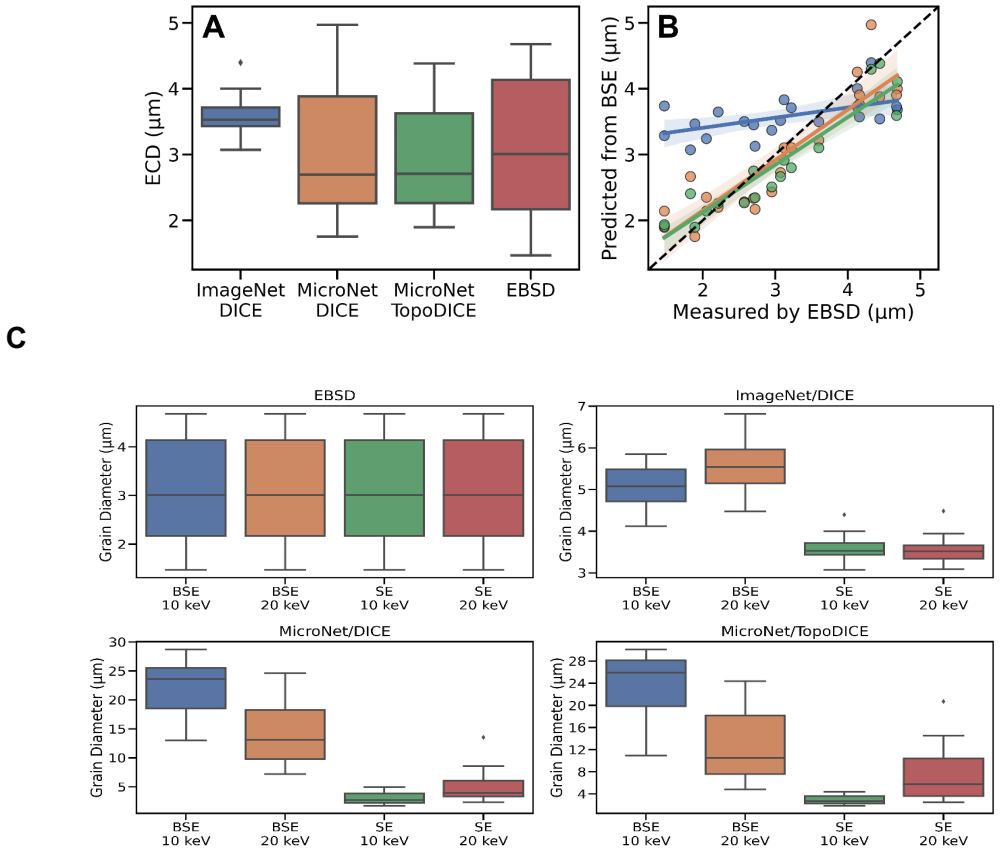


Figure 10. A) Distributions and B) individual ECD predictions on the set of OOD BSE 20 keV images using the SE 10 keV model ensembles, compared against that determined by EBSD. The colors in B match those in A, and the dashed black line shows the ideal fit; C) Distribution of grain diameters predicted from the OOD image set, arranged by model.

Table 12. Statistics for predations of the ECDs of the OOD images, along with corresponding statistics from EBSD measurements taken in the corresponding region.

Training Set	Model	Mean	Std	Min	0.25	0.50	0.75	Max
BSE 10 keV	EBSD	3.06	1.07	1.47	2.17	3.01	4.14	4.68
	ImageNet/DICE	5.05	0.49	4.12	4.72	5.08	5.49	5.85
	MicroNet/DICE	22.47	4.69	12.99	18.53	23.64	25.56	28.72
	MicroNet/TopoDICE	24.07	5.47	10.93	19.84	25.91	28.18	30.10
BSE 20 keV	EBSD	3.06	1.07	1.47	2.17	3.01	4.14	4.68
	ImageNet/DICE	5.54	0.58	4.48	5.15	5.54	5.97	6.82
	MicroNet/DICE	14.71	5.76	7.23	9.80	13.09	18.28	24.66
	MicroNet/TopoDICE	12.65	6.16	4.87	7.62	10.55	18.17	24.39
SE 10 keV	EBSD	3.06	1.07	1.47	2.17	3.01	4.14	4.68
	ImageNet/DICE	3.57	0.30	3.07	3.43	3.53	3.72	4.40
	MicroNet/DICE	2.97	0.91	1.75	2.26	2.70	3.89	4.97
	MicroNet/TopoDICE	2.89	0.82	1.90	2.27	2.71	3.63	4.39
SE 20 keV	EBSD	3.06	1.07	1.47	2.17	3.01	4.14	4.68
	ImageNet/DICE	3.54	0.31	3.09	3.34	3.52	3.66	4.48
	MicroNet/DICE	4.94	2.69	2.38	3.30	3.93	6.06	13.53
	MicroNet/TopoDICE	7.36	4.84	2.49	3.61	5.76	10.45	20.71

Our SEM–EBSD registration approach is similar to that of Shen et al. (2021) who also used correlated SEM–EBSD measurements combined with manual adjustment of the EBSD map. Notably, while their segmentation models trained on this approach could accurately distinguish austenite and martensite phases in dual-phase steel, the models could not determine the exact locations of grain boundaries. This limitation was presumed to arise from the fine, indistinct, and "somewhat fuzzy" boundaries between phases. Despite training on the SE 10 keV image, the models successfully transferred to BSE images taken using different microscopes, settings, and by different operators of samples processed under different conditions with a wider range of mean grain sizes. From this observation, we can conclude that while training requires careful consideration of the physical properties behind data collection, prediction is less sensitive. Training involves learning pixel-to-pixel correlations between the input (SE) and output (EBSD) data, while mean grain size rather than the exact pixel overlap prediction is validates. BSE imaging indeed captures grain boundaries, though at a deeper depth than SE or EBSD. We expect grains at the same location to have similar ECDs across the approximately 250 nm depth captured by BSE compared to the approximately 20 nm depth captured by EBSD.

3.3 Statistical Modeling for Predicting Process-Structure and Process-Property Relationships

Over the years, researchers have dedicated significant effort to understanding the influence of these process variables on the mechanical properties of various metals and alloys (Mishra and Ma 2005, Mishra et al. 2014, Heidarzadeh et al. 2021, Rajamanickam et al. 2009). Yet, the extant literature presently lacks comprehensive, interpretable models to effectively elucidate the intricate relationships between these variables and the resultant properties, particularly with austenitic steel (Mironov et al. 2011, Liu and Nelson 2016, Wang et al. 2014) subjected to FSP. This discrepancy reveals a notable gap in the current research, which tends to heavily rely on black-box models that lack explanatory power.

In today's data-centric world, the field of ML has gained immense traction and ML models have become a common tool for modeling complex systems. However, these models often act as “black boxes,” with their internal workings obscured from the user, making them hard to interpret (Muhamedyev et al. 2020, Petch et al. 2022, Rudin 2019). ML models typically depend on large datasets to deliver reliable predictions, a requirement that is often unmet in most materials science processes, such as FSP. Therefore, the application of ML models can be challenging in these contexts, emphasizing the need for more interpretable models.

Given the aforementioned challenges, this subtask developed both linear and non-linear explainable models that accurately characterize the relationships between the key process variables and the microhardness of FSP 316L SS. The performance of these models was compared to identify the most effective means of capturing these complex relationships. In doing so, this research aspires to provide valuable insights into the intricate interactions between these variables, contributing to the optimization of the FSP and the enhancement of steel properties.

3.3.1 Methods

This task focused on developing relationships between FSP parameters (independent and dependent) and the structure attributes, namely the grain size and the Vickers hardness of the cross-sectional FSP 316L plate samples. The datasets for hardness and grain size analyses included 634 and 114 data points, respectively, collected from 29 distinct samples (SS01-SS34) subjected to 11 different experimental conditions (C01–C11). Table 13 shows the distribution of these observations among samples and their respective experimental conditions.

Table 13. Number of samples, and therefore observations of microhardness and grain size available per condition and sample in the dataset used to develop the PSP relationships in this subtask.

Condition ID	Sample ID	Hardness	Grain Size
C01	SS31	16	5
	SS33	16	3
	SS34	16	5
C02	SS24	16	3
	SS28	12	-
C03	SS01	32	5
	SS23	16	5
	SS09	16	3
C04	SS32	16	5
C05	SS10	16	3
	SS02	90	5
C06	SS03	16	5
	SS11	16	3
	SS21	16	5
C07	SS04	16	5
	SS12	16	3
	SS22	16	5
C08	SS13	16	3

Condition ID	Sample ID	Hardness	Grain Size
	SS05	100	5
	SS17	16	5
C09	SS06	16	5
	SS14	16	3
	SS18	16	5
C10	SS07	16	5
	SS15	16	3
	SS19	16	5
C11	SS08	16	5
	SS16	16	3
	SS20	16	5

It is important to note that, in this study for experimental conditions from C01–C11, four processing parameters were fixed for each case before the experiment. These fixed/set parameters are weld position, tool temperature, traverse velocity, and forge force. The term input features refers to the parameters that were directly controllable or selected prior to performing the FSP experiments. The term target properties refers to the properties governed by the input features, such as mean grain size and average hardness in this subtask. The coordinates of the hardness indentation in the cross-section of the FSP samples were also captured. These coordinates were not part of FSP parameters but might impact predicting hardness since FSP microstructures have varying properties along various locations in the nugget. To study the impact of these parameters on the hardness of FSP samples, we categorized the curated dataset into three groups, as presented in Table 14. In Table 14, dataset 1 only contains the fixed/set parameters, dataset 2 contains the set and dependent parameters, and dataset 3 contains the set parameters and coordinates of hardness indentation. For each dataset, ML models were trained named as Model 1, Model 2 and Model 3 respectively, to study the impact of data type on the predictive accuracy for hardness prediction.

Table 14. Three different datasets for defined for understanding the effects of input variables on the model accuracy for prediction grain size and hardness.

Dataset 1	Dataset 2	Dataset 3
Weld position (mm) Tool temperature (°C) Traverse velocity (mm/min) Forge force (N)	Weld position (mm) Tool temperature (°C) Spindle power (W) Spindle torque (N.m) Spindle speed (rpm) Traverse velocity (mm/min) Traverse force (N) Cross-feed force (N) Forge force (N)	Weld position (mm) Tool temperature (°C) Traverse velocity (mm/min) Forge force (N) X position of the hardness indentation (mm) Y position of the hardness indentation (mm)

3.3.1.1 ML Models

This subtask used traditional ML model frameworks to establish the associations between the input and target variables for the three datasets shown in Table 14. All three datasets were split into train and test sets, where 80% of the data was considered for model training and 20% of

the data was considered as a testing dataset. The train-test splitting was performed in such a way so that there was consistency across the same train and test set from three different data types. To train the ML models, the PyCaret Python library, which simplifies automating ML tasks, including data preprocessing, feature engineering, model selection, and deployment was implemented. We trained 20 different regression ML models with 10-fold cross validation on the training data. MAE was chosen as a performance metric and among all the models developed, the tree-based models, such as Random Forest (RF), Extreme Gradient Boosting, and Extra Tree Regressor, showed the best performance, which is consistent with our previous works for regression type analysis in the materials science domain. To have consistency across the different datasets, Random Forest Model (RFM) was chosen for further analysis, such as feature importance and prediction on the holdout test set.

3.3.1.2 Generalized Linear Models

A linear mixed model (LMM) was used to model the grain size of the cross section in individual FSP samples as a function of the input parameters of interest. An LMM integrates fixed factors such as the main effects of input features or variables and their interactions, alongside random effects to handle variability within and across samples, enhancing the estimation of relationships by minimizing dataset noise. The selection process involves addressing multicollinearity among predictors, where variables with weak correlations ($|r| < 0.7$) are chosen for model building. This comprehensive approach includes constructing a full model with linear, logarithmic, and quadratic terms, evaluated using the corrected Akaike Information Criterion to balance model fit against complexity, particularly in smaller samples. Further refinement is achieved through model averaging, which combines parameter estimates from the best-performing models to assure robustness and generalizability of the results.

3.3.1.3 Symbolic Regression Models

In this subtask, we also employed a symbolic regression approach as a pivotal component of our modeling methodology. Symbolic regression is a powerful computational technique that seeks to discover mathematical expressions that best fit a dataset, to discover underlying relationships and patterns inherent in the data. Unlike conventional regression methods that require the predefinition of a functional form, symbolic regression autonomously evolves and combines mathematical operations and variables to construct equations that capture the complexity of the data. Optimization algorithms that iteratively refine the equations to minimize the difference between model predictions guided this process and observed outcomes. By allowing the exploration of a broad space of mathematical expressions, symbolic regression offers a flexible framework to unveil intricate and non-linear associations between variables.

We used commercially available symbolic regression tools called TuringBot, which is highly engineered and robust, with broad applicability in many domains. The base functions offered by this software are the following:

- Arithmetic: addition, multiplication, division
- Trigonometric: sin, cos, tan, asin, acos, atan
- Exponential: exp, log, log2, sqrt, pow
- Hyperbolic: sinh, cosh, tanh, asinh, acosh, atanh
- Logical: smaller, greater, equal, different, logical_or, logical_and
- History: delay, moving_average

- Other: abs, floor, ceil, round, sign, mod, gamma, erf

3.3.1.4 Approach to Model Development

To train the symbolic regression model, we used an 80%/20% training/testing dataset split. During training, we modeled arithmetic, trigonometric, and exponential bases functions separately.

We initially used a RFM on dataset 1 to develop preliminary correlations and identify the factors of importance. Following that, we used a LMM and symbolic regression to determine the explainable relationships between the parameters of importance identified by the RFM and the target features, namely mean grain size and average microhardness. Finally, we explored the effect of training a dataset on the accuracy of predictions from a data-driven framework by training the ML models on datasets 1–3.

3.3.2 Results and Discussion

In this section, we present the results and discussion on statistical models developed for predicting process-structure and process-property relationships. Specifically, we sought to associate FSP global process conditions with the mean grain size observed in each processed sample, as well as the average Vickers' microhardness of the specimens.

3.3.2.1 Parameter Correlations

It is important to know how the individual input features correlate with each other and how the individual input features correlate with target properties before employing ML models. Correlations will also highlight the multicollinearity in the dataset, if any, which will impact the accuracy of the linear mix model and symbolic regression model. We determined the Pearson Correlation Coefficient (PCC) among the features and feature-target property accordingly. The PCC value ranges from -1 to +1, and the closer the value gets to -1 or +1 means the features are highly correlated. Values close to +1 signify strong positive correlation, and close to -1 signifies strong inverse relation. Figure 11 and Figure 12 shows the PCC matrices and PCC of process parameters and dependent variables for the process parameter—grain size and process parameter—microhardness datasets.

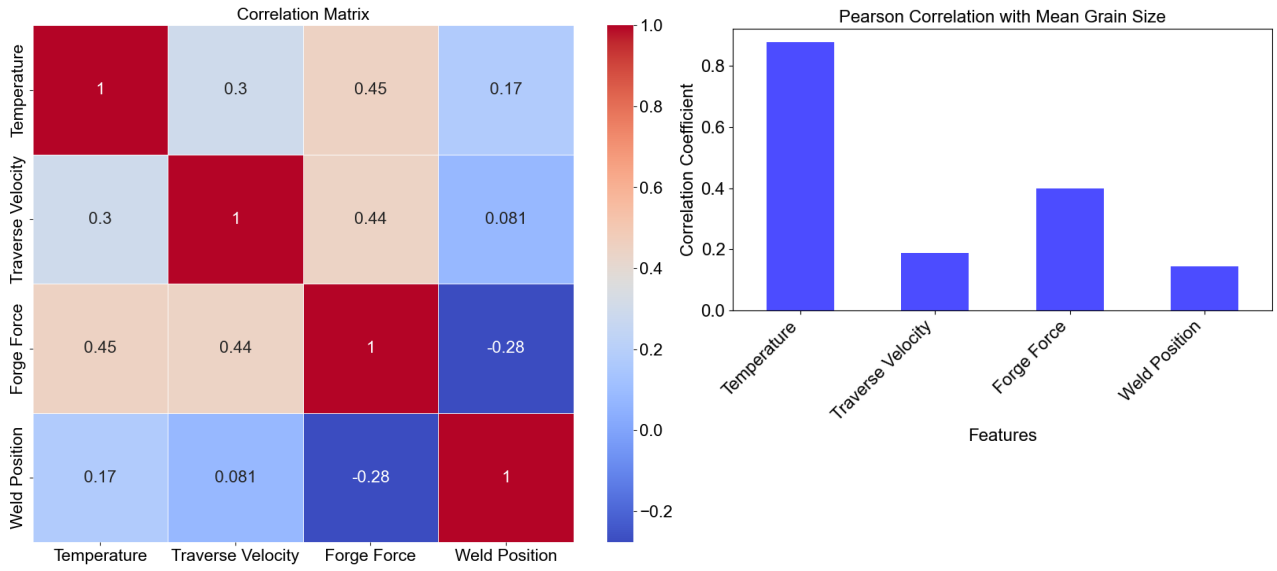


Figure 11a. PCC values for inter-parametric correlations (left), and process parameter-mean grain size correlations (right).

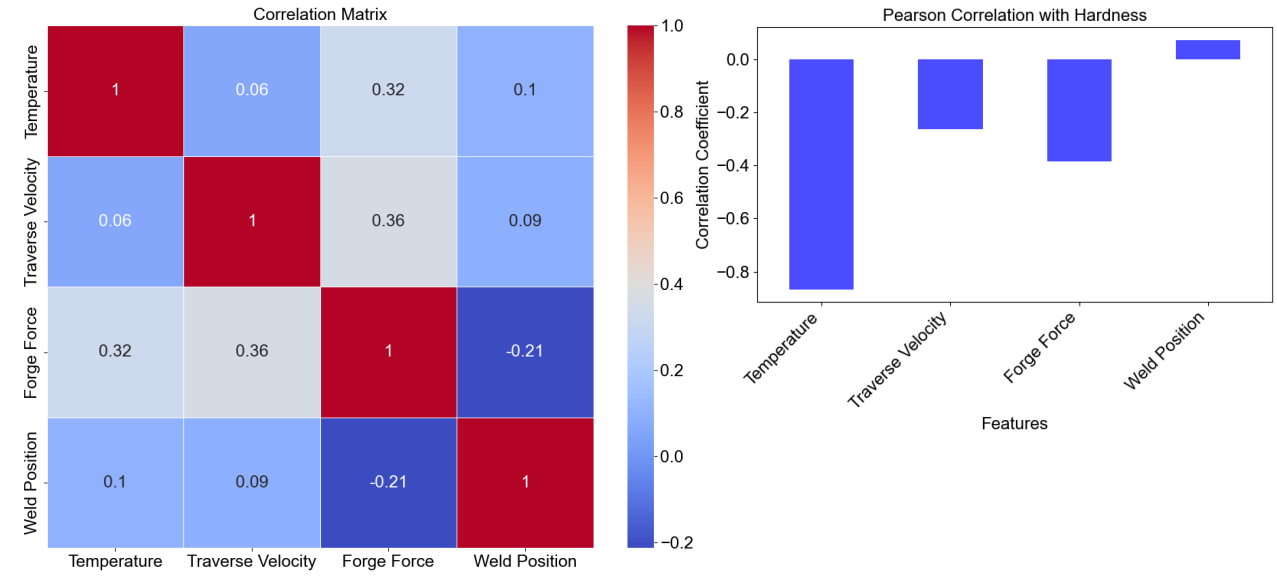


Figure 12. PCC values for inter-parametric correlations (left), and process parameter-mean microhardness correlations (right).

We see the FSP process parameters do not have any strong correlations among themselves. Therefore, all the input features can build explainable models, such as the LMM and symbolic regression model. For the case of tree-based modes like RF, multicollinearity has little impact on model performance (Chen and Guestrin 2016, Haitovsky 1969, Hill and Judge 1987), and therefore, the input features can be used freely to build the RFM. We note that the temperature has a strong negative correlation with hardness and a strong positive correlation with mean grain size. It is important to note that PCC fails to capture the nonlinearity in the dataset. To study the synergic impact of these parameters on the hardness and mean grain size, ML models were employed in this study.

3.3.2.2 Identifying FSP Parameters of Importance

Figure 13 shows the predicted values of microhardness and mean grain size for individual predictions on hold-out test dataset using the RFM developed on the training dataset. The R^2 values for the models predicting microhardness and mean grain size are 0.86 and 0.82, respectively. The model performance increases significantly when the mean predicted values are considered based on experimental conditions, both for hardness and grain size. The mean predictions are almost identical with respect to the actual mean values based on each experimental condition, suggesting that RF is an excellent tool to predict the properties of FSP products just by considering a relevant training dataset.

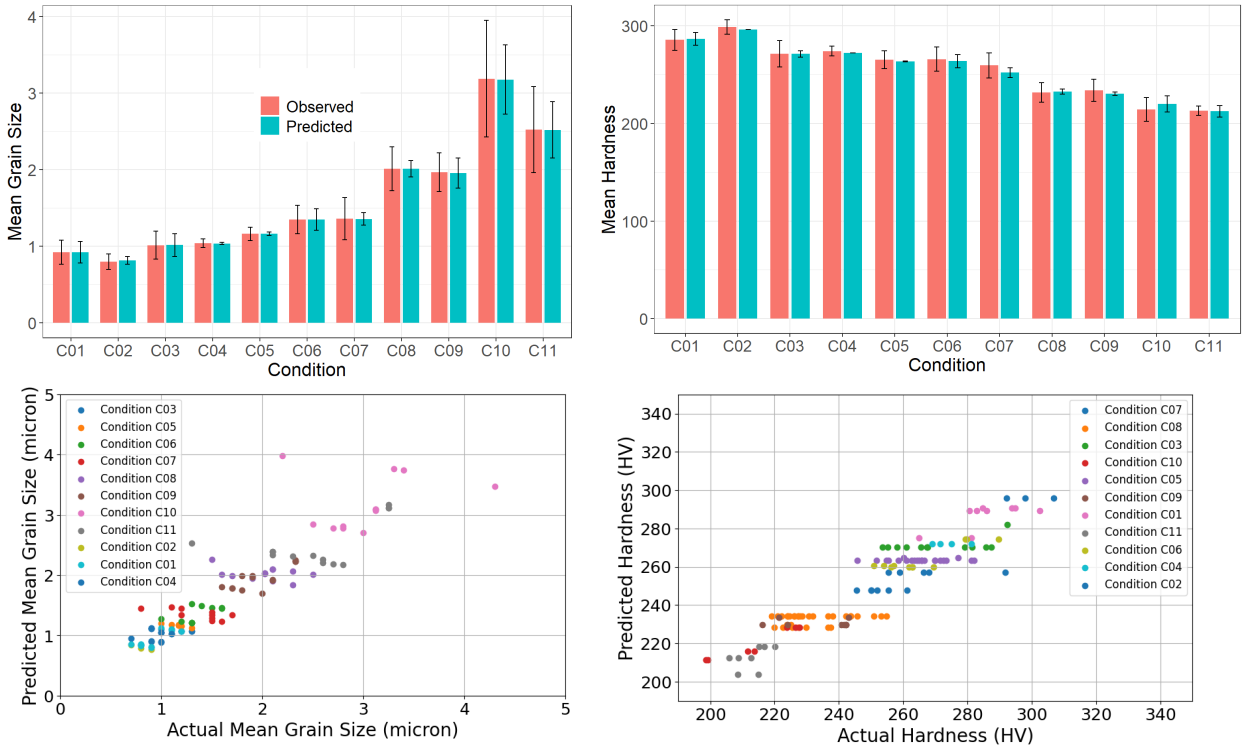


Figure 13. Mean and local grain size and microhardness predicted as a function of FSP process condition for the test data using RFM.

Figure 14 shows the feature importance associated with the RFM for hardness and mean grain size, respectively. Here, feature importance is typically determined by measuring the contribution of each feature to the model's predictive performance. In permutation feature importance, it is measured as the increase in mean squared error (MSE) when the values of a specific feature are randomly shuffled, thus disrupting the relationship between that feature and the target variable. This information is used to rank and select the most influential features, aiding in feature selection and interpretation of the model's behavior. From feature importance, we see that the mean grain size and microhardness prediction models recognize temperature as the most important parameter that significantly affects the mean grain size and hardness of the FSP 316L steel samples. Results show that temperature has a positive impact on mean grain size and a negative impact on mean microhardness, in line with observations from FSP literature (Bahari-Sambran et al. 2024, Chang et al. 2004, Heidarzadeh et al. 2015).

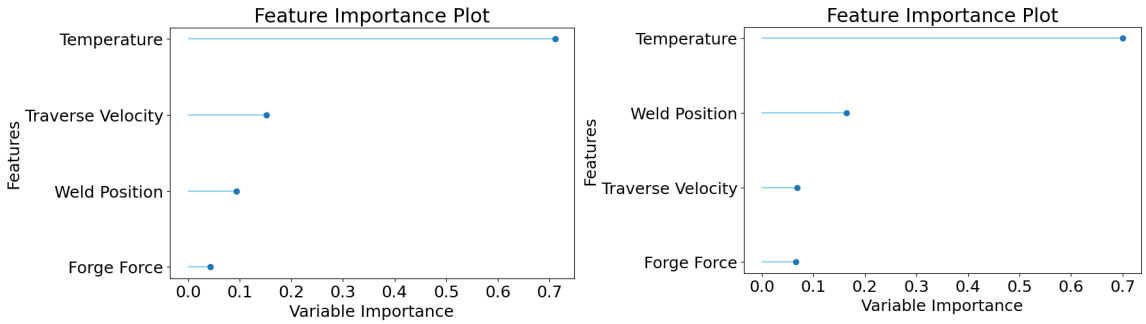


Figure 14. Feature importance for hardness (left) and mean grain size (right) prediction using the RFM.

One limitation of the feature importance generated from the ML model is that it only highlights the score with respect to the importance of the features that govern the target property. Feature importance fails to show the trend of the target property as a function of the input features. To address this, we applied two explainable models namely, LMM and symbolic regression, to the FSP process-structure and process-property datasets. These methods can learn the interactions and trends in the input features and the resulting impact on the target properties.

3.3.2.3 Explainable Models for Understanding Process-Structure and Process-Property Relationships

Figure 15 shows the marginal relationship between temperature and microhardness, as interpreted from the LMM and symbolic regression model, respectively. Accompanied by a 95% confidence band, the plots underscore that an increase in temperature generally leads to a decrease in microhardness when all other variables are held constant. Notably, the symbolic regression has tighter confidence intervals compared to the LMM. LMM shows a clear negative correlation between temperature and hardness, with relatively wide confidence bands reflecting the variability of hardness measurements. In contrast, the symbolic regression model shows consistent and parallel trend lines without depicting confidence intervals, suggesting a more deterministic relationship with tighter confidence bands. The results also elucidate the connection between traverse velocity and microhardness for both models. The representation from the LMM suggests a non-linear relationship, showing that microhardness decreases with increasing traverse velocity up to a certain point and then slightly increases, revealing an optimal range of traverse velocity where hardness reaches its minimum. Conversely, the symbolic regression model shows a pronounced negative relationship between traverse velocity and hardness, particularly at higher weld positions, with tighter confidence bands reflecting more deterministic predictions.

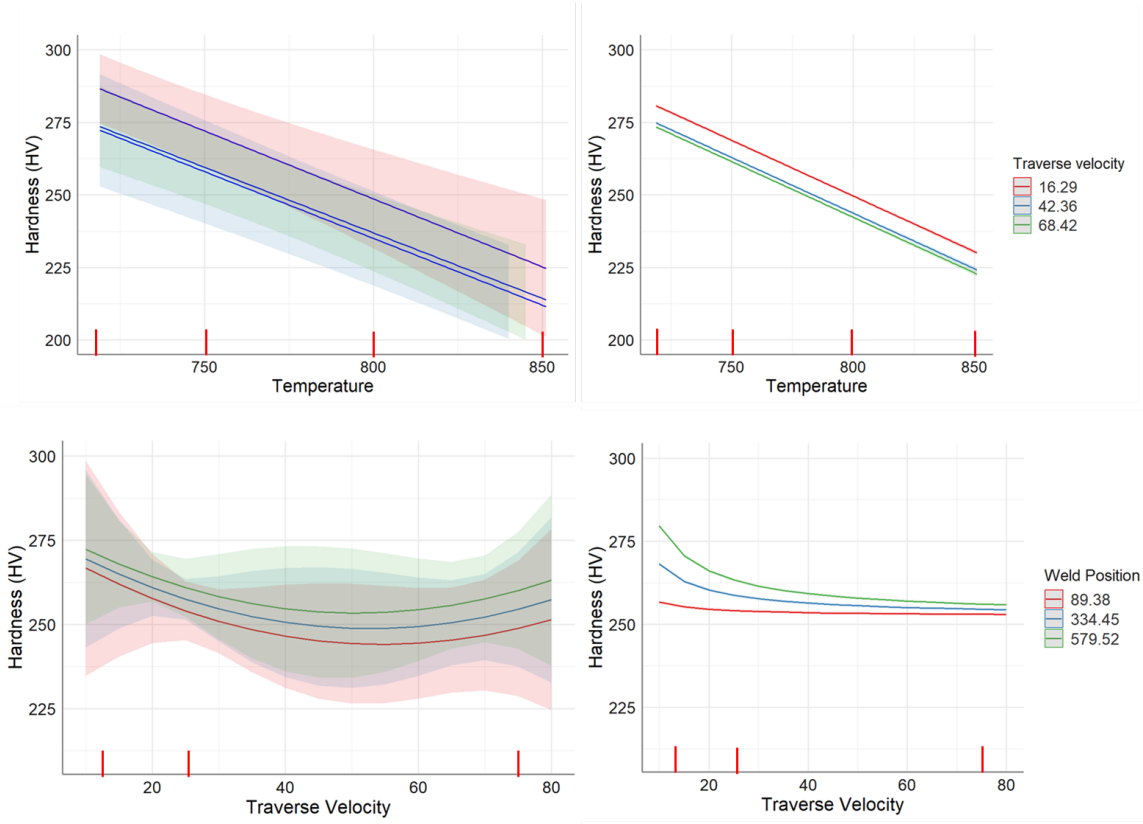


Figure 15. Effect plots from LMM (top and bottom left) and symbolic regression (top and bottom right) showing relationships between temperature and hardness at different fixed traverse velocity values, as well as traverse velocity and hardness at different fixed weld position values, with associated confidence bands. The red marked lines along the x-axis are the experimental data points.

Similarly, Figure 16 depicts the marginal relationship between FSP process parameters and mean grain size, as inferred from the LMM and symbolic regression model, respectively. The plots, along with a 95% confidence band, demonstrate that as temperature increases, there is typically a corresponding increase in the average grain size, assuming all other factors remain constant. Traverse velocity exhibits a slightly positive relationship with grain size. We also note the impact of temperature on the average size of grains at various traverse velocities. LMM shows a positive correlation between temperature and mean grain size. The wide confidence bands represent the variability of the measurements. The symbolic regression model exhibits consistent and parallel trend lines with narrower confidence bands, showing a stronger deterministic relationship between temperature and mean grain size.

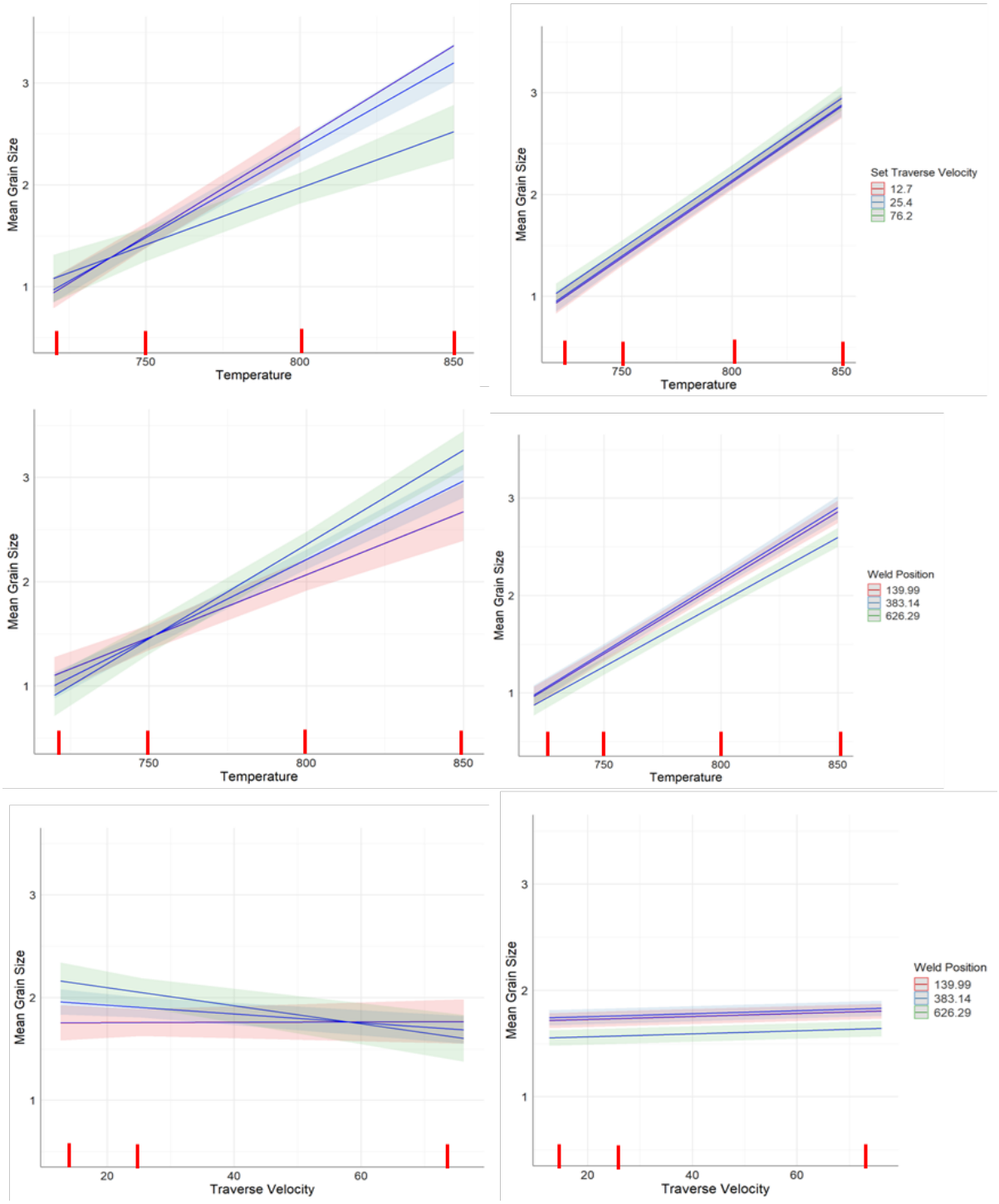


Figure 16. Effect plots showing the relationship between temperature, traverse velocity, weld position, and mean grain size, as interpreted by the LMM and symbolic regression model. Plots on the left are derived from the LMM, whereas those on the right are produced by the symbolic regression model. The red marked lines along x-axis are the experimental data points.

We also investigate the impact of temperature on the average grain size for various weld positions. Both models show a positive correlation between temperature and mean grain size for increasing the weld position owing to the accumulation of thermal energy in the sample. The symbolic regression model once again offers more precise confidence intervals compared to the LMM in this instance as well.

Figure 16 also shows the relationship between traverse velocity and mean grain size for increasing weld position, as predicted by the LMM and symbolic regression models. The LMM shows a negative correlation at the extreme weld positions of 140 mm and 626 mm, while showing a slightly positive correlation for the medium weld position of 383 mm. In contrast, we see a weak positive linear correlation between these parameters in the symbolic regression model, with narrower confidence intervals observed across all weld positions, which is more representative of the experimental findings. While these relationships varies slightly from both models, the flexibility and deterministic nature of the LMM approach make it a preferred option, but predictions from both models align closely, particularly when the 95% confidence bands are taken into account.

3.3.2.4 Effect of Dataset on Model Prediction Accuracy

Figure 17 shows the PCC values for process parameter interaction for datasets 1, 2, and 3, respectively, developed to identify the effect of training factors on model accuracy. The PCC value for a particular pair > 0.7 was considered as a strong correlation in this study. We note that the input features/FSP process parameters do not have any strong correlation in dataset 1. However, we observe a strong correlation among temperature, spindle speed, spindle torque and spindle power, and among traverse velocity, traverse force and cross-feed force for dataset 2.

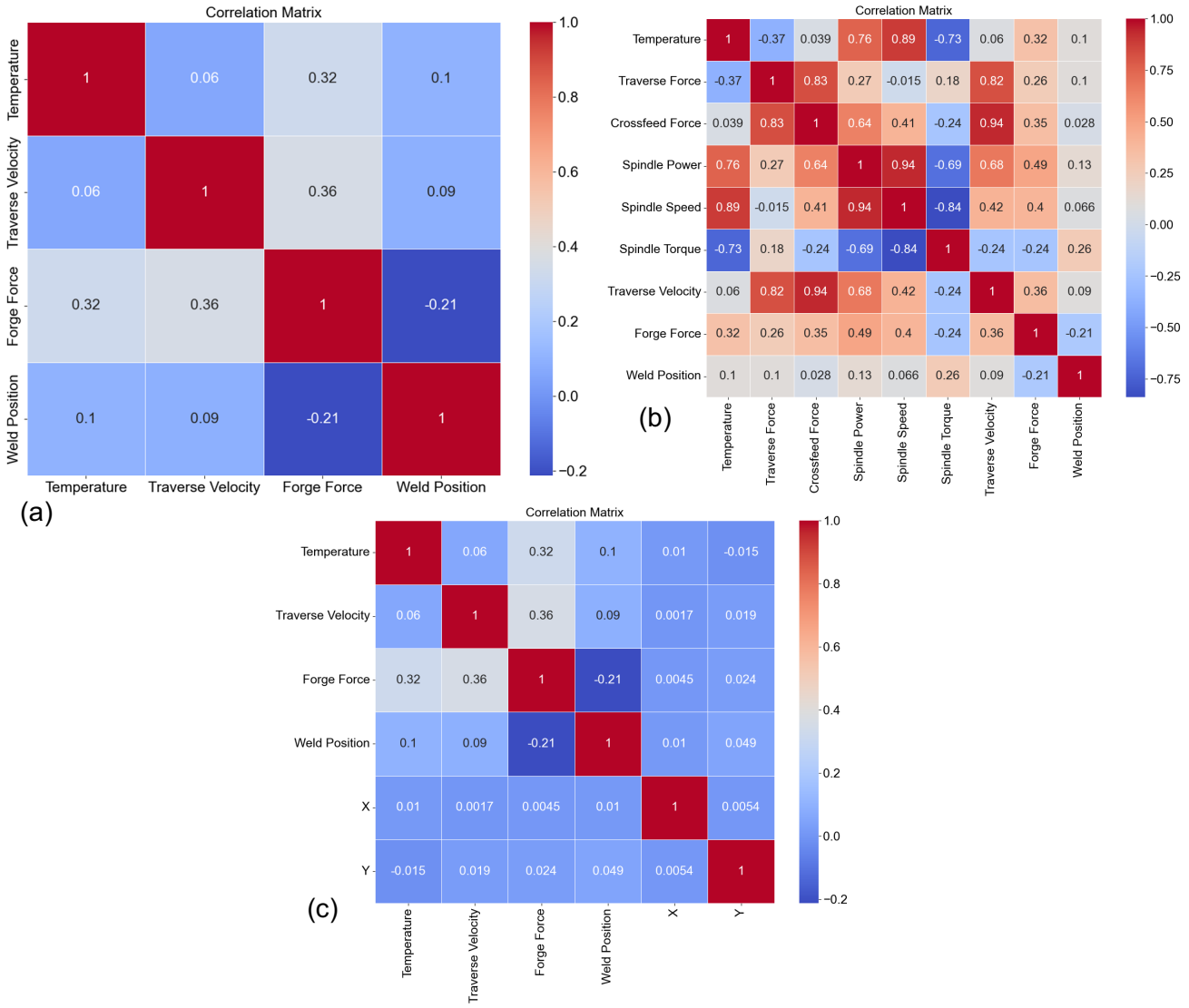


Figure 17. PCC map for FSP process parameter mutual correlations (a) dataset 1, (b) dataset 2, and (c) dataset 3.

Figure 18 shows PCC values for process parameter and hardness interaction for datasets 1, 2, and 3, respectively. Temperature and other correlated parameters, such as spindle speed and spindle power have a strong negative correlation with hardness. In other words, increasing these parameters will decrease hardness in general for FSP 316L plates and within the boundary (upper and lower values) of the process parameters. As noted before, PCC fails to capture the nonlinearity in the dataset and to study the synergic impact of these parameters on the hardness ML models were employed, as discussed in the next section.

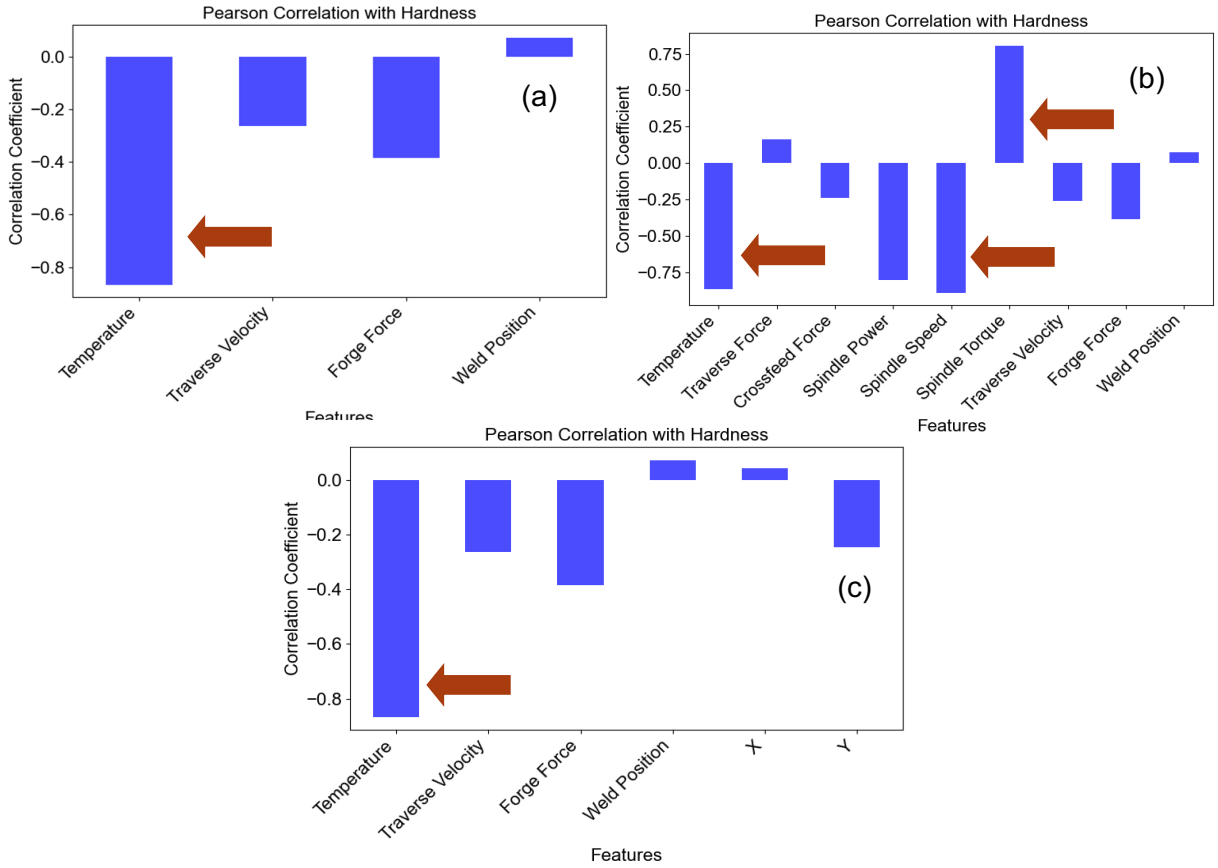


Figure 18. PCC map for FSP process parameter–average microhardness interactions for (a) dataset 1, (b) dataset 2, and (c) dataset 3.

Figure 19 shows the feature importance associated with the RFM for hardness prediction trained on the three different datasets. Here, feature importance is typically calculated by measuring the contribution of each feature to the model's predictive performance. In permutation feature importance, it is measured as the increase in MSE when the values of a specific feature are randomly shuffled, thus disrupting the relationship between that feature (FSP parameter) and the target variable (microhardness). This information is used to rank and select the most influential features, aiding in feature selection and interpretation of the model's behavior. We see that temperature is the most important parameter that controls the hardness of the FSP processes steel samples for dataset 1. Temperature and other parameters related to temperature, such as spindle speed, spindle power, and spindle torque are the top four contributors for dataset 2. Interestingly, the Y location of the hardness indentation is ranked second after temperature in dataset 3. All of these findings align with literature where increasing tool temperature or Y-indentation location (corresponding depth into the plate) can decrease the average microhardness of the processed specimen.

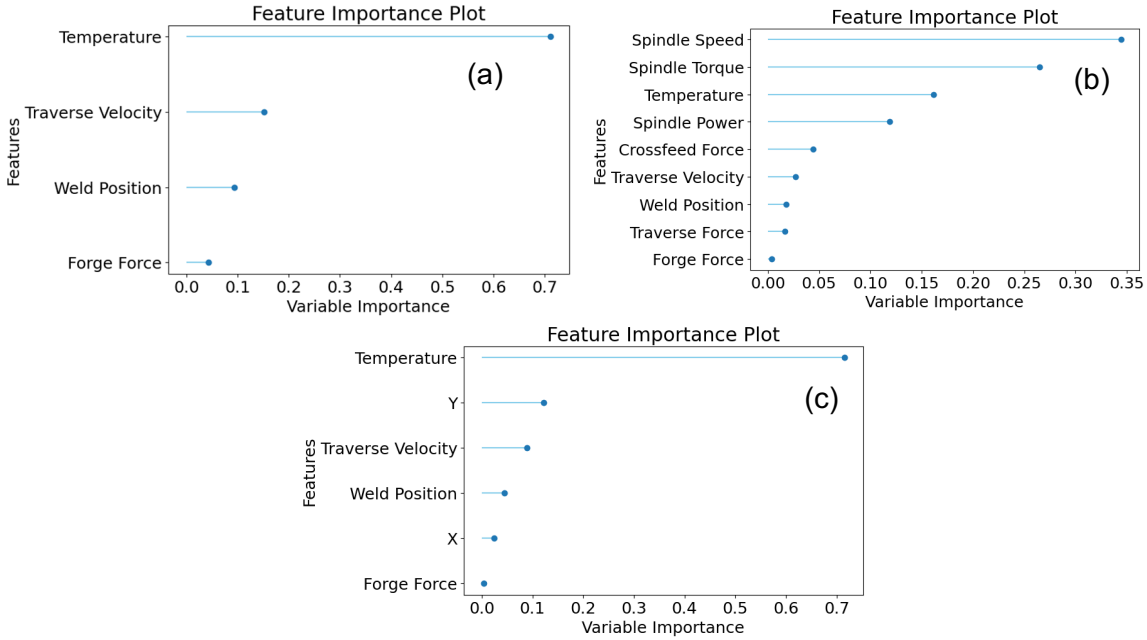


Figure 19. Feature importance determined from RFMs trained on (a) dataset 1, (b) dataset 2, and (c) dataset 3 for predicting average microhardness of FSP 316L plates as a function of different combinations of global or local process parameters.

Figure 20 shows the parity plot for hold-out test data for dataset 1, dataset 2, and dataset 3. Adding the dependent parameters, such as power, increases the R^2 value only by one point compared to dataset 1. However, adding the features based on hardness indentation, such as X and Y coordinates for hardness indentation model performance, increases significantly from R^2 = 0.86 to R^2=0.97 for dataset 3. This is an interesting finding that shows that training on both global and local process parameters, while being effective, may not provide all the information required for accurately predicting the microhardness of FSP 316L as can be obtained when training on the specific cross-sectional locations at which the microhardness indentations were made. While the Y-indentation location is an important factor, as seen from Figure 19 above, it is also clear that the addition of X-location information in the training data has contributed to an improved R^2 value comparatively. The RFM importance factor from dataset 3 also shows that X location is one of the top five factors of importance.

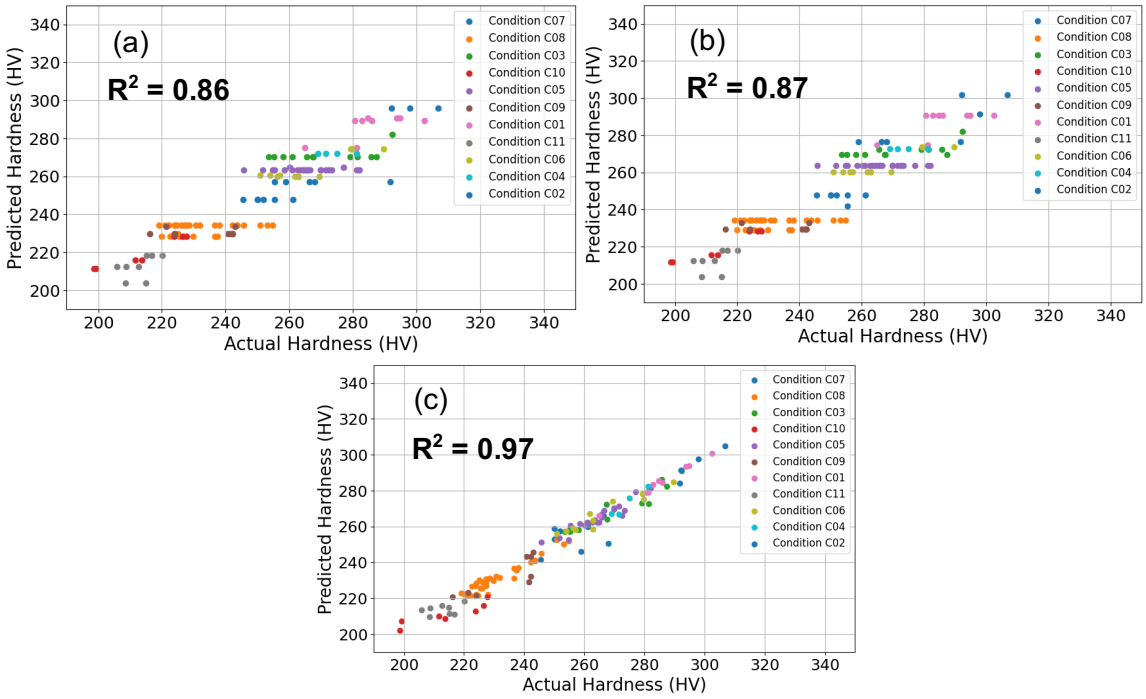


Figure 20. Predicted microhardness as a function of observed microhardness of FSP 316L plates determined by the RFM trained on (a) dataset 1, (b) dataset 2, and (c) dataset 3.

In conclusion, this work presents a comprehensive study on the impact of different types of meta data for FSP on the predictive accuracy for grain size and hardness prediction of 316L SSs. The results suggest that there is no significant difference in R^2 values (0.86~0.87) on the hold-out test dataset for models developed with fixed parameters (dataset 1) and models developed with fixed plus dependent parameters (dataset 2). However, the R^2 value increased significantly (0.86 to 0.97) when geometric coordinates of the hardness indentation are incorporated (dataset 3) in the model. Across all the datasets, temperature and parameters related to temperature, such as spindle speed, spindle torque shows most dominance on hardness prediction. These findings suggest the importance of proper meta data curation and incorporation to build a robust and reliable ML model to predict mechanical properties on hold out unseen datasets. It is important to highlight that the findings obtained from the ML models are valid only for the provided materials and within the confines of the dataset used in this study. It is crucial to note that any alterations to the materials may result in variations in the model outputs. Therefore, caution should be exercised when extrapolating or applying these results to scenarios with different materials.

4.0 Developing Framework for MCPC Control

In this section, we present the approach for developing the CONTROL framework based on which local process parameters or microstructural features may be predicted given the other process parameters whose information is typically collected by the several sensors attached to the FSP and welding equipment. We describe the outcomes from three tasks focusing on predicting temperature as a function of transitory process parameters observed during the initial stages of processing, predicting processed region depth or local tool rotation rate (which can eventually control the processed region/tool temperature) in the steady-state portion of the process, and grain size using ultrasonic transducer responses and local processing conditions. Results from the framework development, namely the model associations demonstrated by the accuracy of the predictions in the subsequent sections, specifically address Objective B.

The central idea of the CONTROL framework is that it is viable to predict the local process parameters and microstructural features using models trained with sufficient information accurately and use the predictions to change the local process parameters applied to a part at a specific time period. The first task describes the approach to predicting local tool temperature based on the power used to manufacture a sample during FSP in the transient stages. The second task focuses on predicting local process parameters in the steady state of processing. The third task develops a framework for microstructural feature prediction through the use of local process parameters used during steady-state FSP and NDE data collected via the ultrasonic testing modality. Important to note is that all three tasks are exploratory and proof-of-principle in nature; they are not mature enough to support real-time instrument control. Future work will consider that question for specific use cases.

4.1 Physics-Informed Neural Networks for Transitory Parameter Prediction

This task seeks to model and control friction-stir welder plunge and dwelling tool temperature using physics-informed ML surrogates of the underlying physical processes.

FSP involves four distinct stages: plunging, dwelling, traversing, and extracting. In the initial plunging stage, the FSP tool makes contact and engages the workpiece. The material plastically deforms, and the internal temperature rapidly rises. The initial plunge and dwell stages are critical for setting the thermomechanical conditions for the subsequent stages. Setting these conditions is critical for weld success: most tool wear occurs during the plunge and dwell stages and unsuccessful temperature control leads to over- and under-shooting of temperature, potentially causing controller stability issues. Well-understood thermomechanical system properties and behaviors are essential for preserving tool life and setting optimal conditions for the traversal stage of the weld.

The multi-scale physics of FSP often necessitate the need for high-fidelity, well-resolved from-first-principles physics simulations for each operating condition for exploration of FSP process parameter space. Performing such simulations on the scale required for heuristic control (i.e., a table lookup for desired operating conditions) is infeasible. ML-based surrogate models can aid in approximating hard to resolve or computationally expensive model components in otherwise well-founded physics-based models.

In the plunge and dwell stages of FSP, power from the FSP machine is transferred to the tool and converted to heat, the measure of which is recorded by a thermocouple over the duration of

the weld. This task seeks to (1) model this temperature profile as a function of process parameters, a system ID problem, and (2) use this model to craft a control input to guide the FSP tool to a desired set temperature suitable for subsequent stages.

4.1.1 Methods

The conversion of machine input power to internal heating of the tool and workpiece is governed by the First Law of Thermodynamics (i.e., all input/output energy must be accounted for). This is true at any observational scale, whether that be at the molecular level or system-wide properties. In the setting of the plunge and dwell stages, we choose to construct a model at the granularity of the observed temperature and machine processing parameters using an energy accounting scheme derived from the First Law of Thermodynamics as a starting point.

A time-dependent energy accounting scheme can be framed as a first-order differential equation for energy: $\frac{dE}{dt} = \dot{Q} - \dot{W}$, where E is energy, $\dot{Q}$ is the rate of heat input into the system, and $\dot{W}$ is the work done by the system on its surroundings. Given that we are interested in how power is converted into a temperature (with no interactions with the surroundings), this can be re-written in a more useful form that relates temperature dynamics with input and output heating terms for a lumped volume with capacitance C; $\frac{dT}{dt} = \frac{1}{C}(\dot{Q_{in}} - \dot{Q_{out}}) = \frac{1}{C}\left(\eta P(t) - h(T - T_{sink})\right)$. Here, we have assumed that temperature dynamics are directly related to the difference in heating through a power input $P(t)$ scaled by some unknown conversion factor η and convective cooling. While we have neglected many physical effects in this minimum viable model for temperature dynamics, the essential physics are present, and the overall behavior can be directly linked back to from-first-principles physics. In this model, the conversion factor from power to internal heating is unknown, as is the capacitance of the lumped volume and the heat transfer coefficient h.

The data science task is to tune these unknowns so that the model best reproduces experimental data. Recent advances in ML for dynamical systems have enabled gradient-based optimization for systems that obey differential equation relationships. Termed *Neural ODEs* (Chen et al. 2018), this class of methods allows for automatically differentiating ordinary differential equations -based model output with respect to inputs and parameters, enabling DL integration with physical models. Here, we replace the conversion factor η with a neural network $\eta(T, P; \theta)$: $\mathbb{R}^2 \rightarrow (0,1)$ that relates tool temperature and input power to a conversion efficiency. For this work, the neural network is a simple feed-forward multilayer perceptron with a single hidden layer of size 15 and a sigmoid activation function. The output layer is similarly sigmoid activated to constrain the output to be in the range (0,1). Fitting this model to available data constitutes the system ID problem.

Given a differentiable system ID model, model weights can be frozen and a trainable network inserted as a stand-in for exogenous inputs. In this case, the exogenous input is the controllable FSP machine power profile, which can be tuned to guide the tool temperature to a desired set temperature or temperature-time profile.

4.1.2 Results and Discussion

Full results can be seen in (Koch et al. 2024). In summary, the differential equation-based model, when imbued with first-principles physics and the adaptivity of the neural network term, excels at relating power input to temperature dynamics. The model can reproduce data with high accuracy, including in applications to unseen data and experimental conditions. The model

is interpretable and robust by construction and guaranteed to obey the First Law of Thermodynamics. No additional model training, regularization, or post-hoc analysis is required to enforce these physics.

This research culminated in a proof-of-concept result for control of the temperature over time by crafting a power profile using the trained system ID model. Given a desired set temperature (700 Celsius in this work), a neural network was inserted as a stand-in for $P(t)$, such that the model could learn the power profile to reach the desired set temperature. This style of control is referred to as open-loop control since there is no feedback provided to the control algorithm. Experimental validation of this control strategy is reserved for future work.

4.2 Recurrent Neural Networks for Steady-State Parameter Prediction

In this task, we examined the ability of DL models to predict the in-process parameters during FSP to aid in the design of experiments. Manufacturing processes need predictive and computationally efficient methods to better develop targeted materials and products. Historically, physics-based simulations have permitted scientists to explore process parameters and various other configurations, enabling them to avoid running experiments that are likely to yield poor results. Such methods are typically computationally expensive and constrained to idealized states. Advances in ML, specifically DL, have enabled the creation of surrogate models can learn from experimental data, acquiring statistical relations for many process parameters. These models can often run in a fraction of the time that their counterpart, physics-base models require. By now, DL has been applied to a wide range of problems in materials science, from the classification of materials based on performance, to the synthesis of molecules, to the simulation of material dynamics. More specifically to FSP, there has been a large amount of work examining prediction of defects and control fracture failure through the image domain of surface appearance (Sudhagar et al. 2019, Verma et al. 2021) and with relation to process parameters (Verma et al. 2021). Closer to our method and goal, prior work has deployed DL networks to monitor force represented over time to classify defects (Boldsaikhan et al. 2011).

We focused on modeling process parameters, namely tool rotation rate or spindle speed and processed depth as a function of process temperatures, forces in all the axes, and tool traverse rate. Having the ability to predict processed depth and tool rotation rate provides insight into the defects that might arise during processing and the amount of strain a workpiece volume experiences. We note that the processed depth or tool traverse rate at a specific time is a function of the process conditions previously experienced by the workpiece, as well as the material properties of the workpiece. Most ML models take a collection of input x and convert it into a predictive function $f(x)$ via supervised training. For many tasks, the true value, y, can be a category or real number. The goal is to optimize f such that we minimize the error between $f(x) = \hat{y}$ and y.

The problem being tackled in the current study looks not at a single point prediction but changing dynamics over time. A sequence to sequence (seq2seq) model is a common learning paradigm designed for such situations. A seq2seq model attempts to learn a mapping from one sequence of points to another. A classic problem in ML that fits this situation is machine translation, where we have an input sentence and need to output the words in another language. For a domain such as depth prediction, the depth value at t depends not only on prior operating values of x_t ($x_{t-1} \ldots x_{t-n}$), but also on previous predictions $\widehat{y_{t-1}} \ldots \widehat{y_{t-n}}$. Thus, a seq2seq model benefits from the ability to model the state over time with memory. One family of

architectures capable of performing the seq2seq task are recurrent neural networks (RNNs) that store such memory states and are flexible in their ability to handle time windows of varying size. We use a subclass of such models called long-short-memory networks (LSTM) (Hochreiter and Schmidhuber 1997) based on their ability to model longer sequences. Closely related works include, DeepTemp (Wight et al. 2022), which uses a RNN to model temperature as a function of time-given process parameters such as tool traverse rate and tool rotation rate during the manufacturing of aluminum alloy 7075 tubes using ShAPE™. Other work has addressed modeling temperature evolution in FSP, framing the problem as a universal differential equation (Koch et al. 2024) and parameterized using a neural network.

4.2.1 Methods

Our goal in this subtask was to predict the tool rotation rate and processed depth at a given time as a function of the FSP process parameters. We note in this instance of how depth is measured on the friction-stir welder that the uncertainty of measurement is high. In this task, we examine the ability of DL models to predict the in-process parameters during FSP to aid in process control as needed. Manufacturing processes need predictive and computationally efficient methods to better maintain the process conditions in predefined ranges and adapt in an agile fashion to the variations seen in location owing to several external factors. Historically, physics-based simulations have enabled exploration of process parameters and avoided running experiments that are likely to yield poor results. Such methods are typically computationally expensive and constrained to idealized states. Advances in DL have enabled the creation of surrogate models that can learn from experimental data, acquiring statistical relations for many process parameters. These models can often run in a fraction of the time that their counterpart, physics-base models require.

In this work, we evaluate two LSTM models (Hochreiter and Schmiduber 1997), one based on a time stepper that outputs processed depth or tool rotation rate evolution over a window of time, and the second based on a residual time stepper that predicts the rate of change of the processed depth and rotation rate from the initial conditions. We compare the performance of these LSTMs against two baselines, one that uses the start value as a constant (baseline) and a single-step model that operates only on the prior output (or start condition), the previous processing variables, and the current processing variables ($f(\widehat{y_{t-1}}, x_{t-1}, x_t) = \widehat{y_t}$. We explore the relationship between the number of parameters, the complexity of the model, and the performance.

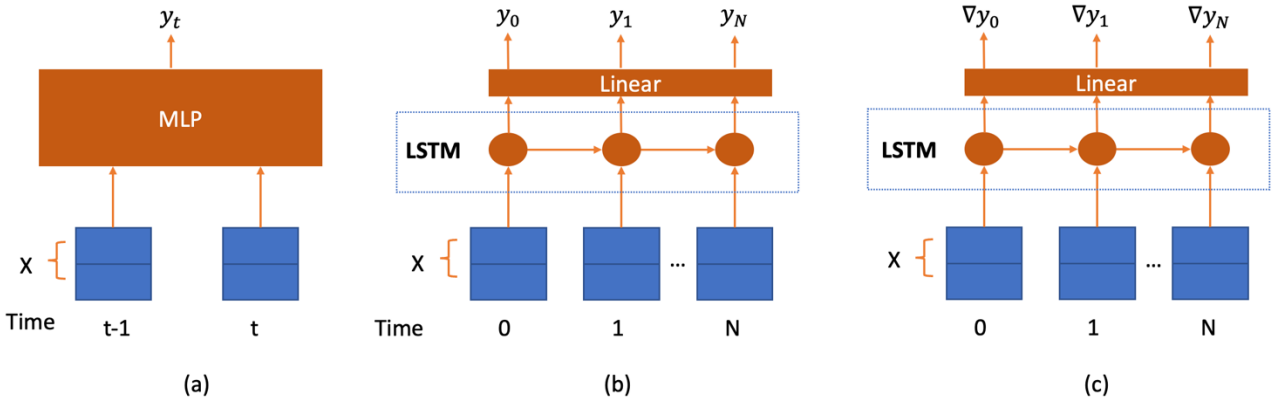


Figure 21. The modeling paradigms we examine a) is a single-step history model that uses a multilayer perceptron (MLP), b) a time-stepper LSTM model that outputs the actual value, and c) is a residual time-stepper LSTM model that outputs the rate of change from the initial condition.

We first split experiments into train and test sets between the ten runs with different configurations. We choose to train on nine experiments and test on the tenth one. This method represents real-world use cases where we will apply the model to a new experimental process window that has the desired tool rotation rate and processed depth.

We used two different LSTM, which are a type of RNNs for modeling the relationship between the process parameter plots and the dependent variables, namely tool rotation rate and processed depth. Our hypothesis is that the memory-based models will enable better performance over time-window based models. The first method consisted of using a time-stepper method where, given a previous depth estimate y_0 we attempted to predict $y_{1 \ldots N}$, where *N* is the time duration size. We represent our LSTM model as a function $f(x_{0\ldots1}, y_0) = \widehat{y_{1\ldots N}}$. The memory or hidden state here allows the model to consider the dynamics of the process parameters and target variables over time. The second method we employed was a residual variation of the LSTM model, where the model outputs a rate of change from a base condition across time steps $f(x_{0\ldots1}, y_0) = \Delta\widehat{y_t}$. We compared the proposed LSTM performance and two baselines against experimental data measuring the error across time windows. The first baseline uses the start value of the target variable as a constant across all time steps. This simple constant baseline is to show that models are using this information more than just using this constant provided. The second baseline predicts using a single-step model that predicts the target variables based on the starting condition, the process parameters used in the previous time step, and the process parameters available in the current time step.$f(\widehat{y_{t-1}}, x_{t-1}, x_t) = \widehat{y_t}$. This model represents a test of the inductive bias for the memory component that our LSTM models optimize for past states. This model represents a test of the inductive bias that memory of past states provides our LSTM models versus the ability to model process parameters locally with limited time steps.

We scaled the LSTM models' complexity through both the hidden state size (width) and the number of layers (depth). We used MSE as our training loss function and the Adam optimizer (Kingma and Ba 2014) with a learning rate of 0.001 for 1000 epochs for all models. We ran our evaluation across all combinations of our ten experimental runs, with one experiment being held out and tested on while being replaced in the train set. We preprocessed the data by normalizing all variables individually. We only performed normalization on the nine experimental run data in the train set and applied those statistics by feature to the test set. We determined the MAE across each experiment with the variance for all folds. Next, we measured the

relationships between the internal width and depth of the LSTM's dimension. Motivated by the fact the choice of how deep and wide to make a model varies per domain and dataset, we experiment with various model design choices. For longer sequential processing, width has been found to be more important than depth. We measure the LSTM model across 64 and 1024 hidden state dimensions and 2 to 3 layers. We used 1024-time steps for our base models that correspond to ~51 s. The approximate relationship between the time steps used in the model and the experimental time window is because of the heterogeneous time between samples. We finally examined the scaling efficacy of the LSTMs when training and testing on time steps corresponding to 2048 (102 s) and 4096 (204 s). To accommodate this increase in modeling length, we used our largest model with 3 layers and a hidden state dimension of 1024.

4.2.2 Results and Discussion

Table 15 presents the MAE with variance calculated for all folds for the LSTMs and the baseline models for comparison as a function of the time window. Results show that the MAE of the LSTM residual time-stepper model was the least, while that of the LSTM time-stepper model was also comparably low. In fact, both the models' MAE did not show statistically significant distinction. The highest MAE was seen for the baseline model with a constant start as anticipated. The results shown in Table 15 comparing the methods of modeling verified our two hypotheses. Our baseline method that uses the start value given to the other models as a constant has the highest MAE and consequently, the lowest performance. When looking at the model-based approaches, the MLP memory-less model had only 23% of the performance we observed from the time-stepper LSTM model. This shows that giving the models access to additional prior state information can enable the model to find more accurate relationships among the process parameters across time compared to when local time step only information is provided. However, time-stepper methods depend on the ability of the model to output the actual value. We observe when modifying this to be residual, there is a 13% increase in performance. This change is in line with literature (Legaard et al. 2023) stating this allows models to learn dynamics rather than value ranges plus dynamics.

Table 15. Performance across various model framing methods on tool depth and MAE prediction.

Model	Memory Size	Method	Predicted Tool Depth ± MAE (µm)
Baseline	0	Start constant	10.2 ± 3.04
MLP	1	Time-stepper	7.8 ± 1.96
LSTM	1024	Time-stepper	1.9 ± 0.83
LSTM	1024	Residual Time-stepper	1.6 ± 0.41

Figure 22 shows two examples of predicted depth of FSP tool penetration as a function of the time period of processing. The negative values specifically show instances of when the model predictions align well with the experimental data (ground truth) and also when there is a marked deviation, capturing the distribution in MAE. In both instances, we note that the LSTM shows a closer alignment with the experimentally recorded depth over the other models.

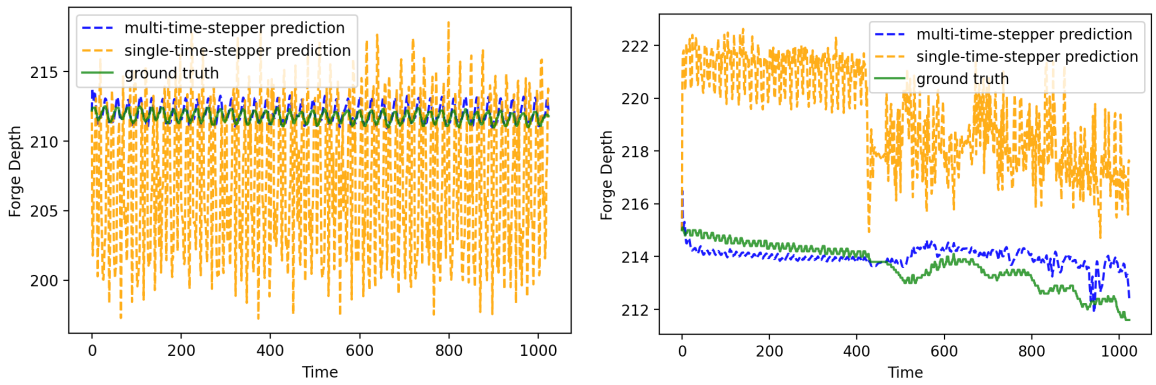


Figure 22. An example of different model predictions plotted over 1024 time points, roughly 51 seconds. We see visually that the proposed multi-time-stepper model (blue) more closely models the true ground truth depth (green) pattern compared to the single-time stepper model (orange) (left), and a failure case of both models not capturing the depth behavior over time. Our multi-time-stepper model (blue) has sporadic jumps and does not capture the true ground truth depth (green) (right).

Table 16 shows the variation of MAE of the time-stepper models as a function of the hidden state dimensions ranging from 32 to 1024 and for 2 and 3 model layers. The least MAE was observed for two layers and a hidden dimension of 128. For both the layers explored, the highest MAE and variance of 9.12 ± 0.45 μm were seen for the 1024 hidden dimension. When measuring the complexity of the models with respect to performance, we found that increasing depth resulted in higher performance ranges compared to increasing the depth only. We find more consistent results independent of the hidden state dimension with the larger depth models. Our best performing model, taking into account the variance, is the LSTM with 3 layers and a hidden state dimension of 64. This exercise showcases the need for validation of methods and the careful hyperparameter tuning that can be done to increase performance.

Table 16. LSTM predicted tool depth and MAE for different model layers and hidden state dimensions.

Layers	Hidden State Dimension	Predicted Tool Depth ± MAE (μm)
2	32	1.8 ± 1.42
2	64	1.5 ± 0.41
2	128	1.0 ± 0.44
2	265	1.1 ± 0.41
2	1024	1.9 ± 2.61
3	32	1.8 ± 1.32
3	64	1.1 ± 0.35
3	128	1.12 ± 0.40
3	265	1.13 ± 0.76
3	1024	9.12 ± 0.45

Table 17 shows the MAE of the time-stepper LSTM as a function of time step (and corresponding experimental process time), where we examine the model's ability to predict larger time sequences. We see the performance of the model decreasing as the time-step size

increases. The MAE is 19 times higher when doubling the experimental processing time from size from ~51 to ~102 seconds. We see an approximately double that increase when going from ~102 to ~204 seconds. There are multiple factors that could be influencing this performance of the LSTM model. Modeling larger contexts can be more difficult for the model given only the context of the initial processed depth. Secondly, when we increase the time step, we increase the length of training data provided to the model, which decreases the number of unique sequences available for model training. This exacerbates the already critical small data problem in this study and probably leads to higher MAE for the higher time-step cases. Based on these results, such models are best used in smaller prediction windows in this auto-regressive manner.

Table 17. LSTM predicted tool depth and MAE for across various time lengths.

Time Step	Time (seconds)	Performance (MAE in microns)
1024	~51	1.2 ± 0.45
2048	~102	24.9 ± 7.31
4096	~204	44.6 ± 48.20

We next examined the ability of DL models to predict the tool rotation rate (spindle speed) given other model parameters, namely the local FSP conditions. We evaluated the ability of various modeling methods, namely time stepper, residual time stepper, single time stepper, single residual time stepper, and a baseline in predicting the local spindle speed during processing at a time step. For the time-stepper and residual time-stepper analysis, we used an LSTM with 2 layers and 128 hidden dimension size. For the single time-stepper models, we used an MLP with 3 layers and a hidden dimension size of 128. Table 18 shows the MAE obtained from the various models as a function of the memory size used. Our results show our models are doing more than reusing initial conditions when comparing the baseline. We found that the residual model performed worse than the exact value time-stepper model for predicting this value. Similar to modeling depth, we found that the models with memory for the entire sequence have higher performance. However, our MLP time-stepper model achieved stronger performance compared to tool depth modeling.

Table 18. Performance across various model framing methods on spindle speed prediction.

Model	Memory Size	Method	Tool Rotation Rate ± MAE (rpm)
Baseline	0	Start Constant	50.8 ± 14.89
MLP	1	Time-stepper	10.5 ± 8.59
LSTM	1024	Time-stepper	6.9± 3.75
LSTM	1024	Residual Time-stepper	11.2 ± 4.30

Figure 23 shows the tool rotation rate predicted by the single- and multi-time-stepper models and the experimental data (ground truth) recorded during the FSP experiment for two different time windows. For the time window on the left, we note that the prediction models show a similar variation with the experimental data but are capturing the variation closely. For the instance shown on the right, we can see the multi-time stepper (blue) model having a bias toward stability, not capturing the faster rate of change. Here, the single time stepper (orange) has a higher variation across predictions in the wrong direction.

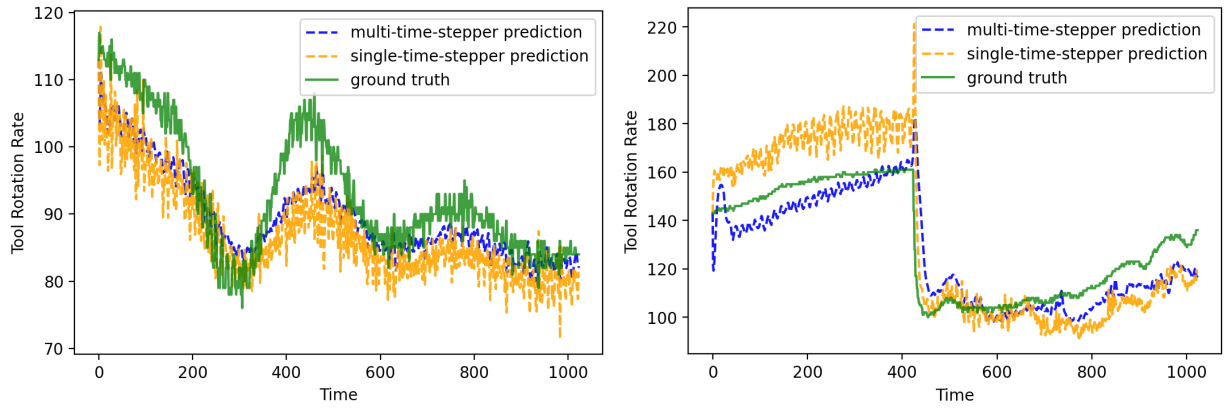


Figure 23. Tool rotation rate predicted by the single time stepper (orange), and multi time stepper (blue) capturing general spindle speed behavior (green) on an unseen experiment (left) along with the experimental local tool rotation rate (ground truth); and failure case for spindle speed modeling where both proposed methods fail to model the initial speed variation (right).

4.3 Grain Size Prediction using Ultrasonic Testing Data

Ultrasonic testing (UT) NDE enables process validation of FSP material and provides a cost-saving method for material characterization. UT is one of many modern NDE techniques applied in manufacturing settings (Wang et al. 2020). If UT NDE signals can be accurately mapped to known microstructure characteristics, a second step of cost reduction could be achieved by data-driven prediction of UT NDE signals using only in-line process parameters, laying the foundation for in-line process control with online defect identification or mitigation.

ML is commonly used for prediction or classification of weld quality and microstructural features based on UT NDE signals (Sun et al. 2023) or directly from process variables (e.g., Hartl et al. 2021). Friction-stir applications of UT NDE have primarily focused on defect detection (Tabatabaeipour et al. 2016) and bulk property characterization (Hynes et al. 2014) in the post-processing stage. In-line fault detection for FSP is rare in the literature (Saxena et al. 2022), but is an important extension of the work presented in this case study. Case studies outlining exploratory data analysis, model construction and evaluation, and recommendations for future work combining ML, UT NDE, and FSP can lay an important foundation for improving the use of ML in UT NDE prediction for FSP performance.

In this work, we contribute one such case study of the prediction of UT NDE signals using in-line process parameters coupled with an extension to microstructural property prediction (grain size). Exploratory data analysis and in-depth visualization techniques are used to explore UT NDE features. ML and statistical techniques are compared to model, predict, and interpret the relationship between in-line process conditions, UT NDE signals, and microstructural properties. Prediction uncertainty is quantified with statistical tools. A variety of general extensions are discussed for hybrid statistical-ML-DL material prediction and process control.

4.3.1 Methods

In this subtask, we developed a framework associating local FSP parameters with mean average voltage obtained from UT NDE performed on the sections of the processed 316L plates under steady-state conditions. An Imaginant PureView-La1 ultrasonic pulser and receiver

module (S/N JE0111) was used to excite the ultrasonic transducer to generate ultrasound into and receive/condition returned signals from a sample. The conditioned signals (waveforms) output by the receiver were subsequently digitized at 500 MHz sampling frequency. The dataset comprises 76.5M voltage measurements corresponding to 51 index locations along the length of the sample, 250 scan positions at every index location, and 6000 sample locations along the depth of the plate. The dimensions of the UT data are visualized in Figure 24.

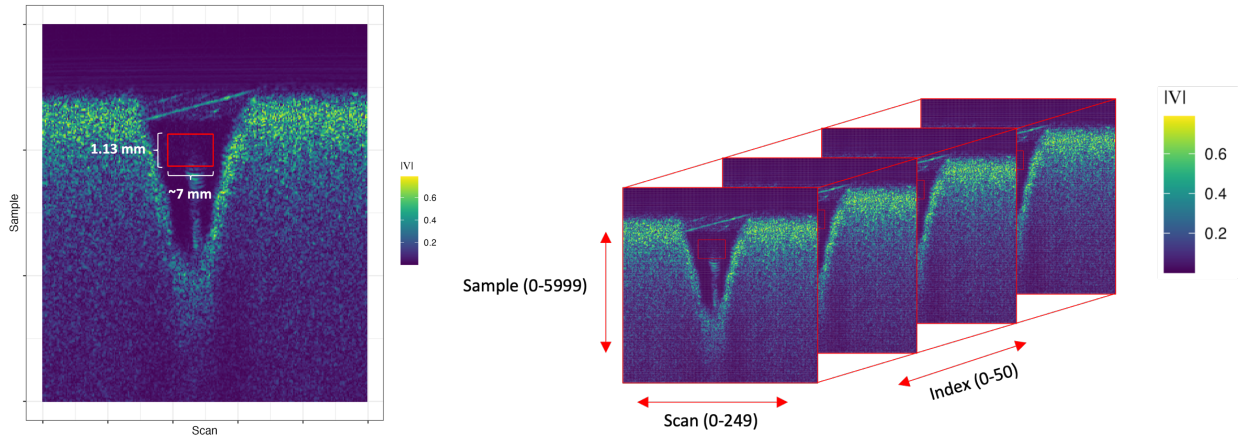


Figure 24. Absolute voltage as a function of sample depth (y-axis) and scan position (x-axis) for a given index location (left); schematic of the UT data demonstrating the difference between sample depth, scan position, and index location. The region bounded by the red box in the image on the left shows the region of interest considered in this effort.

A specific goal of the analysis was to consider sub-regions of the samples where the materials have been thoroughly processed and represent the nugget region. We focus only on data in the bounding box region, marked by red rectangles in Figure 24. In addition to focusing only on the bounding box regions, we also excluded data from indices at the beginning and end of the specimen to ensure the data under consideration is free from effects caused by the specimen edges. The dependent variable, namely the average grain size of the processed 316L steel in the region of consideration, was the ECD provided by the AztecCrystal software package included with the Oxford Instrument's EBSD instrument used for characterizing the microstructure. The response UT signal from each sample (depth of wave penetration) and scan was averaged over the indices to obtain a MAV distribution over an entire NDE specimen, which was produced with a specific set of global and local process parameters. Line plots were obtained from these signals by averaging the absolute voltage over the scan position and sample position, resulting in a single MAV per index, as shown in Figure 25.

Local differences within the specimen can be observed where voltage spikes (MCPC 0118, first row, fourth column) or voltage decreases (MCPC 0083, fourth row, second column). Global differences are seen in the magnitude of MAV differences between conditions. Specimens processed according to Condition C11 have much higher voltage than those processed according to Condition C02. A control specimen is also plotted in the bottom right of Figure 25 with a free-scale y-axis, demonstrating that the unprocessed specimen has the highest overall voltage among all the samples analyzed.

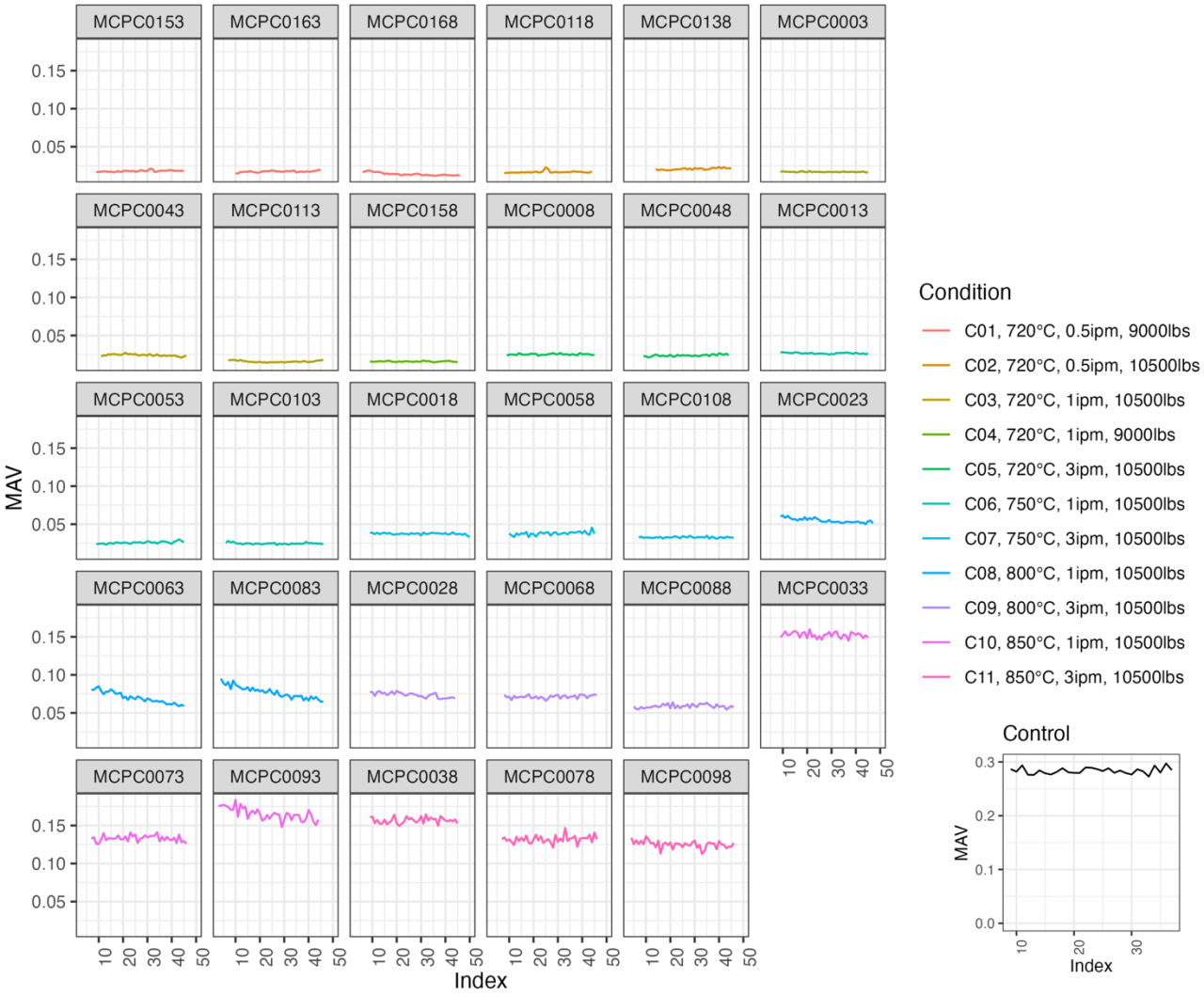


Figure 25. Absolute voltage from UT of FSP 316L SS plates averaged across scan and sample of the NDE specimen as a function of the index location (along the traverse axis of the specimen).

It is possible to visualize the process parameter time series data in addition to exploring the specimen-level UT data. We focus on a small subset of parameters with known importance. Figure 26 plots the local process parameters averaged over index per specimen, namely spindle speed or tool rotation rate, torque, power, and tool temperature against the specimen-level MAV. The process parameters appear to be highly correlated with MAV and each other. This is known and expected, as certain process parameters control other variables (e.g., spindle speed controls tool temperature).

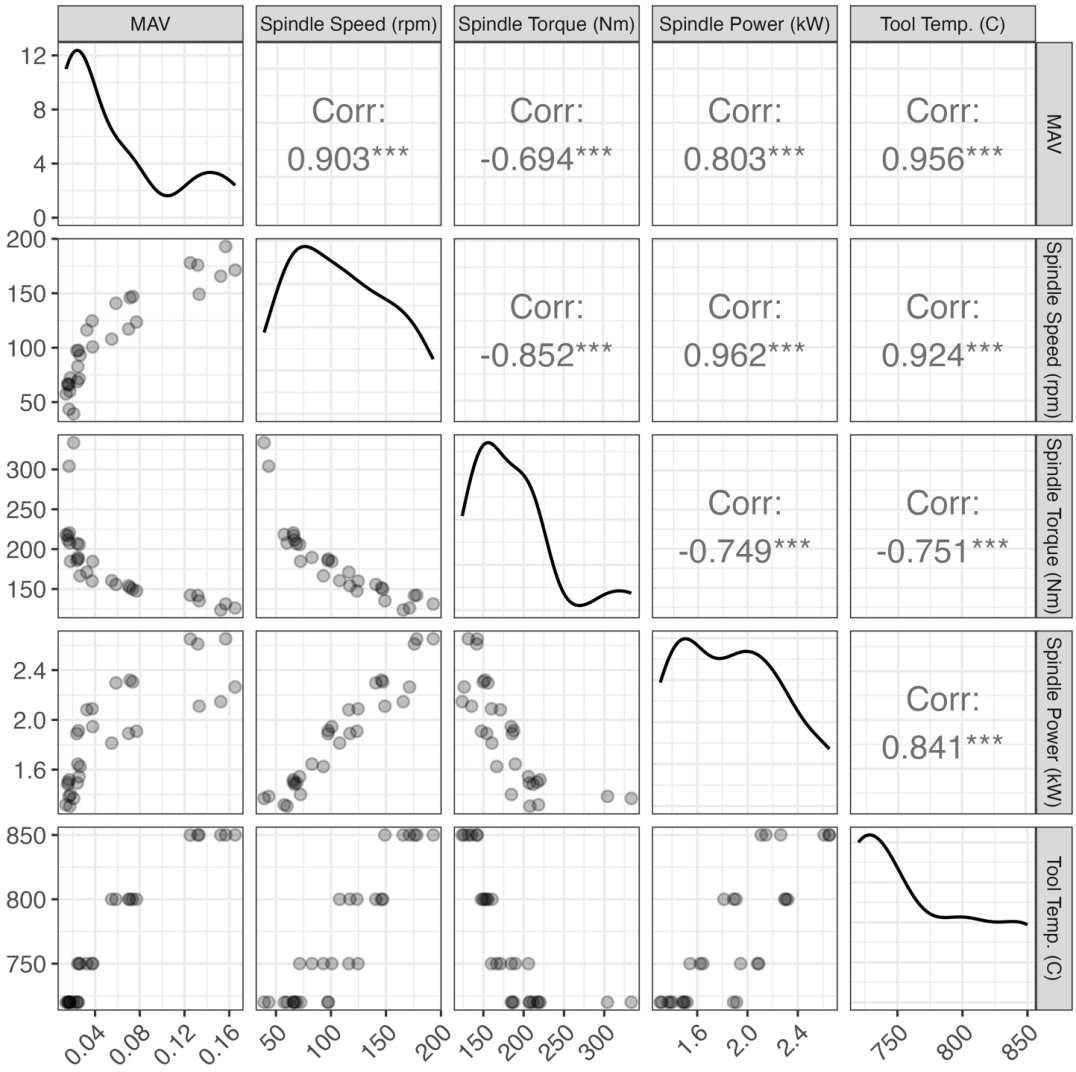


Figure 26. Subset of process parameters, averaged over index and plotted against voltage averaged over sample, scan, and index.

The objective of this task is predicting MAV as a function of process parameters and extending the index-level UT predictions to specimen-level grain size prediction. In our work, we consider a general model shown below in Equation (3):

$$MAV_{i,r} = f(\boldsymbol{x}_{i,r}) + \epsilon_{i,r}, i = n_{r0}, \ldots, n_r, r = 1, \ldots, m \quad (3)$$

where $MAV_{i,r}$ represents the MAV at a particular index, i, of a particular specimen, r, averaged over sample and scan. Here, the vector of scalar process parameter values at a particular index, obtained by averaging the process parameter time series values in a window around each index point is represented by $x_{i,r}$ as Equation (4):

$$\boldsymbol{x}_{i,r} = [x_{i,r,1}, \ldots, x_{i,r,p}]' \quad (4)$$

and $\epsilon_{i,r}$ is the observation error. The indices are bound by the first index, n_{r0}, and the last, n_r.

Statistical and ML methods considered for this prediction task are outlined in Table 3, along with brief descriptions, hyperparameters tuned, and model references. The tidymodels (Kuhn and Wickham 2020) and tidyverse (Wickham 2019) ecosystems were used for preprocessing, model training, hyperparameter tuning, cross validation, test set prediction, and interpretation.

Table 19 Models and hyperparameters used for associating process parameters and MAV in this task.

Name	Preprocessing	Hyperparameters	Model Reference
RF	none	mtry trees min_n	(Breiman 2001), (Kuhn and Johnson 2013)6/10/25 9:24:00 AM
XGBoost	none	tree_depth trees learn_rate	(Chen and Guestrin 2016), (Kuhn and Johnson 2013)
Multilayer perceptron via Keras (MLP)	normalization	hidden_units penalty epochs	(Kuhn and Johnson 2013)
Elastic Net (simple)	none	penalty mixture	(Hastie et al. 2015), (Kuhn and Johnson 2013)
Elastic Net (polynomial)	quadratic + interaction terms	penalty mixture	
Elastic Net (PCA)	normalization principal component analysis	penalty mixture	

In addition to directly fitting Elastic Net (a mixture of lasso and ridge regression) on the data (simple), two additional variations of Elastic Net were considered: (1) The polynomial Elastic Net, which included quadratic terms and all two-way interactions and (2) the principal component analysis (PCA) Elastic Net, which replaces the predictors with five principal components estimated from the normalized polynomial Elastic Net terms.

A stratified 75%-25% train-test split was targeted by grouping each specimen by index quartile and randomly sampling 75% of indices (rounded) per quartile to add to the training set. The rounding resulted in a 78%-22% train-test split. This scheme guaranteed that all index regions of all specimens are present in both the training and testing sets. This is a reasonable decision for model prediction and inference for the objectives of the task. We note applications focused on process control would benefit from holding out entire MCPC specimens or entire regions of specimen indices. Hyperparameter tuning was performed with a 5-fold cross validation with 2 repeats. Resampling methods like cross validation enable performance evaluation without referencing the test/validation set. One rule of thumb is to use 10-fold cross validation with 5 repeats (Kuhn and Johnson 2020). Our choice of 5 folds and 2 repeats sought balance prediction performance with a computational burden. Where target metrics (such as prediction accuracy) suffer, exploring different choices of folds and repeats (or other resampling methods such as Monte Carlo cross validation or bootstrapping) can be prudent.

Prediction intervals (PIs) are crucial for applied predictive modeling. A PI describes the minimum and maximum expected value for an unobserved predicted point with some confidence. PIs are wider than confidence intervals, as they are predicting the limits for a single

point rather than an average. After establishing the RF model as the preferred method, PIs were computed based on estimating the error distribution with the out-of-bag (OOB) error (Zhang et al. 2020). The RF model is an ensemble method where hundreds of regression trees are fitted on a subset of variables and training observations. The individual predictions are averaged to obtain overall predictions. Because a subset of the training observations is used to grow each tree, the withheld OOB observations can be used as a test set. The prediction error of these OOB observations can estimate an error distribution, from which quantiles can be obtained to define upper and lower PIs.

Previously, specimen-level grain size was obtained as an extremely low-dimensional representation of specimen-level microstructure (one value per specimen). After obtaining a model, f, to predict $MAV_{i,r}$, it is of interest to investigate the relationship between specimen-level MAV and specimen-level grain size. The observed specimen-level training MAV is used to fit a statistical model between grain size and MAV. Then, we use this model to map our predicted MAV to grain size predictions.

Formally, we consider the model,

$$gs_r = h(MAV_r) + \epsilon_r, r = 1, \dots, m, \tag{5}$$

where gs_r corresponds to the specimen-level grain size and MAV_r corresponds to the MAV at the specimen level denoted by r. We use the average of index-level averages in the training set. Thus, we can write:

$$MAV_r = \frac{1}{n_{train}} \sum_{i \in train} MAV_{i,r}. \tag{6}$$

In theory, we may be interested in the effect of defining MAV_r as the average of training set index-level averages (as opposed to the average of all raw data, across both train-test sets and without index-level averaging). This effect is negligible for our analysis but may be important in other contexts. After h is estimated, grain size can be predicted by plugging in the average of the index-level test set predictions. That is,

$$\widehat{gs}_r = h\left(\widehat{MAV_r}\right), r = 1, \dots, m, \tag{7}$$

where $\widehat{MAV_r} = \frac{1}{n_{test}} \sum_{i \in test} \widehat{MAV}_{i,r}$. PIs for $\widehat{gs}_r$ can be obtained with a multi-step Bonferroni correction scheme (for more information on the Bonferroni correction, see Bland and Altman 1995):

1. Select a target overall confidence level, $1 - \alpha$
2. Compute PIs for $\widehat{MAV_r}$ based on f to obtain $(\widehat{MAV}_{r,L}, \widehat{MAV}_{r,U})$ at confidence level $1 - \frac{\alpha}{2}$. $(\widehat{MAV}_{r,L}, \widehat{MAV}_{r,U})$ correspond to the lowest and highest values one would expect the true MAV_r to fall between at a confidence level of $1 - \frac{\alpha}{2}$.
3. Obtain grain size predictions for these PIs by plugging them into h. That is, obtain
   a. $\widehat{gs}_r^L = h\left(\widehat{MAV}_{r,L}\right)$

b. $\widehat{gs}_r^U = h\left(\widehat{MAV}_{r,U}\right)$

4. Compute PIs for $\widehat{gs}_r^L$ and $\widehat{gs}_r^U$ based on h to obtain $(\widehat{gs}_{r,L}^L, \widehat{gs}_{r,U}^L)$ and $(\widehat{gs}_{r,L}^U, \widehat{gs}_{r,U}^U)$ at confidence level $1 - \frac{\alpha}{2}$. Retain the lower bound of the prediction based on the lower bound and the upper bound of the prediction based on the upper bound to obtain a new PI pair, $(\widehat{gs}_{r,L}^L, \widehat{gs}_{r,U}^U)$ as the final grain size PI.

Because two PIs were constructed (one for $\widehat{MAV}_r$ in Step 2 and one for $\widehat{gs}_r$ in Step 4) with a confidence level of $1 - \frac{\alpha}{2}$, the overall confidence level is $1 - \alpha$ by the Bonferroni correction. While the new PI pair $(\widehat{gs}_{r,L}^L, \widehat{gs}_{r,U}^U)$ is constructed from two different PIs estimated in Step 4, further correction is not needed as in both cases only one tail was required, effectively resulting in a confidence level of $1 - \frac{\alpha}{4}$ for each bound.

For example, suppose we have an initial confidence level of $1 - \alpha = 0.9$, with $\alpha = 0.1$. The PIs from Step 2 are constructed such that $\Pr\left(MAV_r > \widehat{MAV}_{r,L} \,\&\, MAV_r < \widehat{MAV}_{r,U}\right) = .95$ (the probability that the true specimen-level MAV is between the PIs is equal to 95%). The PIs from Step 4 are constructed such that $\Pr\left(gs_r^L > \widehat{gs}_{r,L}^L \,\&\, gs_r^L < \widehat{gs}_{r,U}^L\right) = .95$ and $\Pr\left(gs_r^U > \widehat{gs}_{r,L}^U \,\&\, gs_r^U < \widehat{gs}_{r,U}^U\right) = .95$. However, we only require that $gs_r^L > \widehat{gs}_{r,L}^L$ and $gs_r^U < \widehat{gs}_{r,U}^U$. Assuming symmetry, $\Pr\left(gs_r^L > \widehat{gs}_{r,L}^L\right) = .975$ and $\Pr\left(gs_r^U < \widehat{gs}_{r,U}^U\right) = .975$, resulting in an overall confidence level of 0.95. Thus, each set of PIs has a confidence level of 0.95, resulting in an overall confidence level of 0.9.

PIs for Step 2 are obtained by a bootstrapping method. For a specimen, the sample of index-level test set predictions is bootstrapped and averaged 10,000 times. These bootstrapped specimen-level predictions are used to obtain a distribution of the absolute error between the bootstrapped specimen-level predictions and the true specimen-level MAV based on the test set. The $1 - \frac{\alpha}{2}$ percentile of this distribution of absolute errors is added and subtracted to the test set specimen level prediction to obtain the $1 - \frac{\alpha}{2}$ intervals in Step 2.

4.3.2 Results and Discussion

For each model, the set of hyperparameters resulting in the lowest root mean squared error during cross validation was selected, and the resulting metrics are plotted in Figure 27. The XGBoost and RF performed the best across all metrics. XGBoost has slightly better training performance in terms of MAE, root mean squared error, and R^2, while RF has improved performance in terms of mean absolute percentage error. RF was selected as the model of choice for further consideration, as (1) RF is simple and (2) RF has a straightforward PI computation based on the OOB test set discussed in the approach to model development.

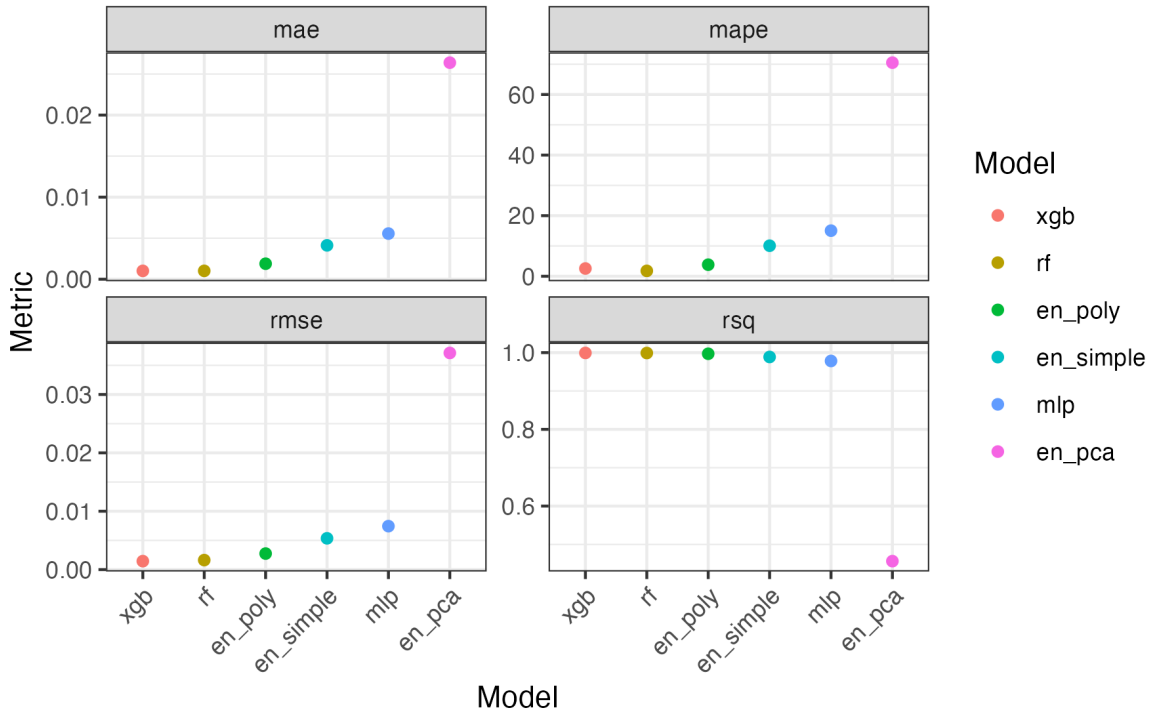


Figure 27. Regression metrics of models with the best performing hyperparameter configuration.

An importance plot in Figure 28 based on the RF model ranks the process parameters in terms of their impact on MAV prediction for a specimen. Tool temperature and spindle torque have the greatest impact on prediction performance, followed by spindle speed and forge position. Index location (traverse location or x-axis location) is less important than most process parameters. This finding does not suggest that location is an unimportant variable *within* conditions or specimens. We anticipate that were a model to be fit only on data from Condition C08, we would expect index location to be an extremely strong predictor.

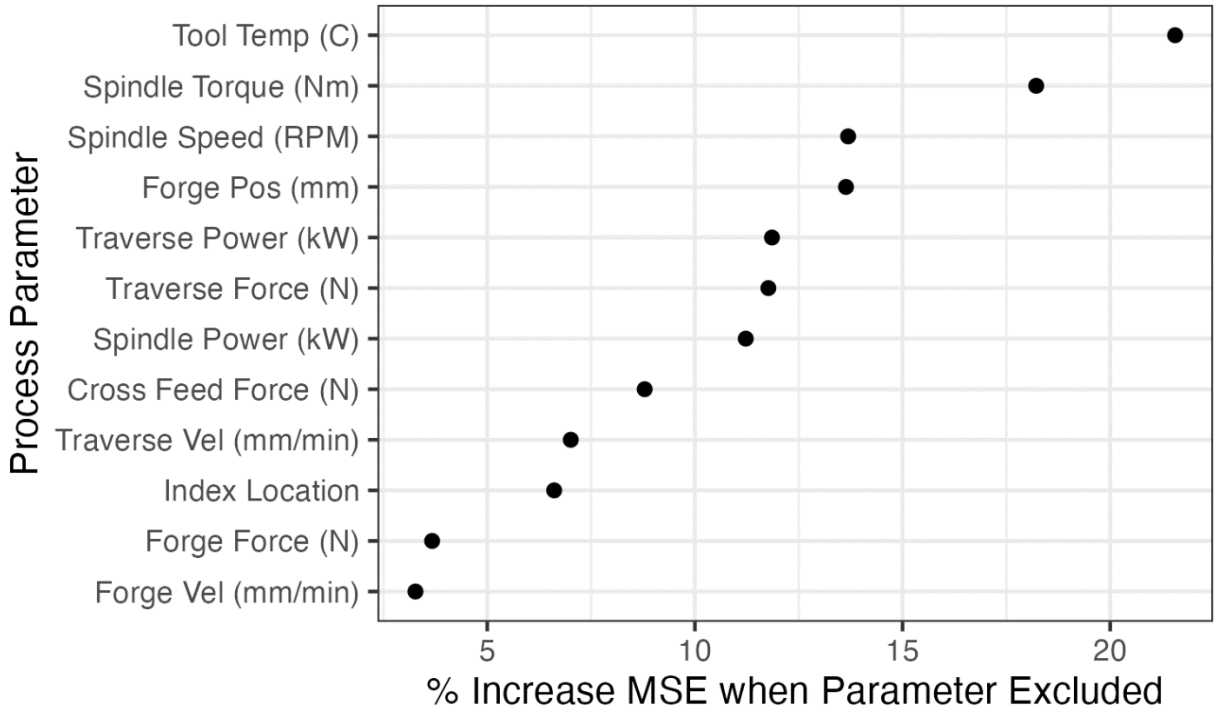


Figure 28. Tool temperature and spindle torque are the most important variables in predicting MAV.

Figure 29 plots the test set observations versus predictions, emphasizing that the RF performs well on test data prediction across all conditions. Prediction variability increases for the higher conditions, and this is expected as the overall magnitude of MAV is higher under higher conditions. Figure 29 also plots the mean and standard deviations of the training observations, testing observations, and testing predictions, grouped by condition. This plot shows that (1) the train and test sets had similar mean and variance, and (2) the RF predictions on the test set are similar to the experimental data. The prediction error can be visualized by plotting the test set predictions along with their 90% PIs over indices, faceted by the specimen. Note that in this plot the y-axis (MAV) is not standardized across panels. The closed circles with intervals correspond to predictions and their PIs, while black asterisks correspond to test set observations. While in some cases, the observations fall outside the PIs, the predictions follow local trends, such as the decreasing behavior in MCPC 0083 (first column, fourth row).

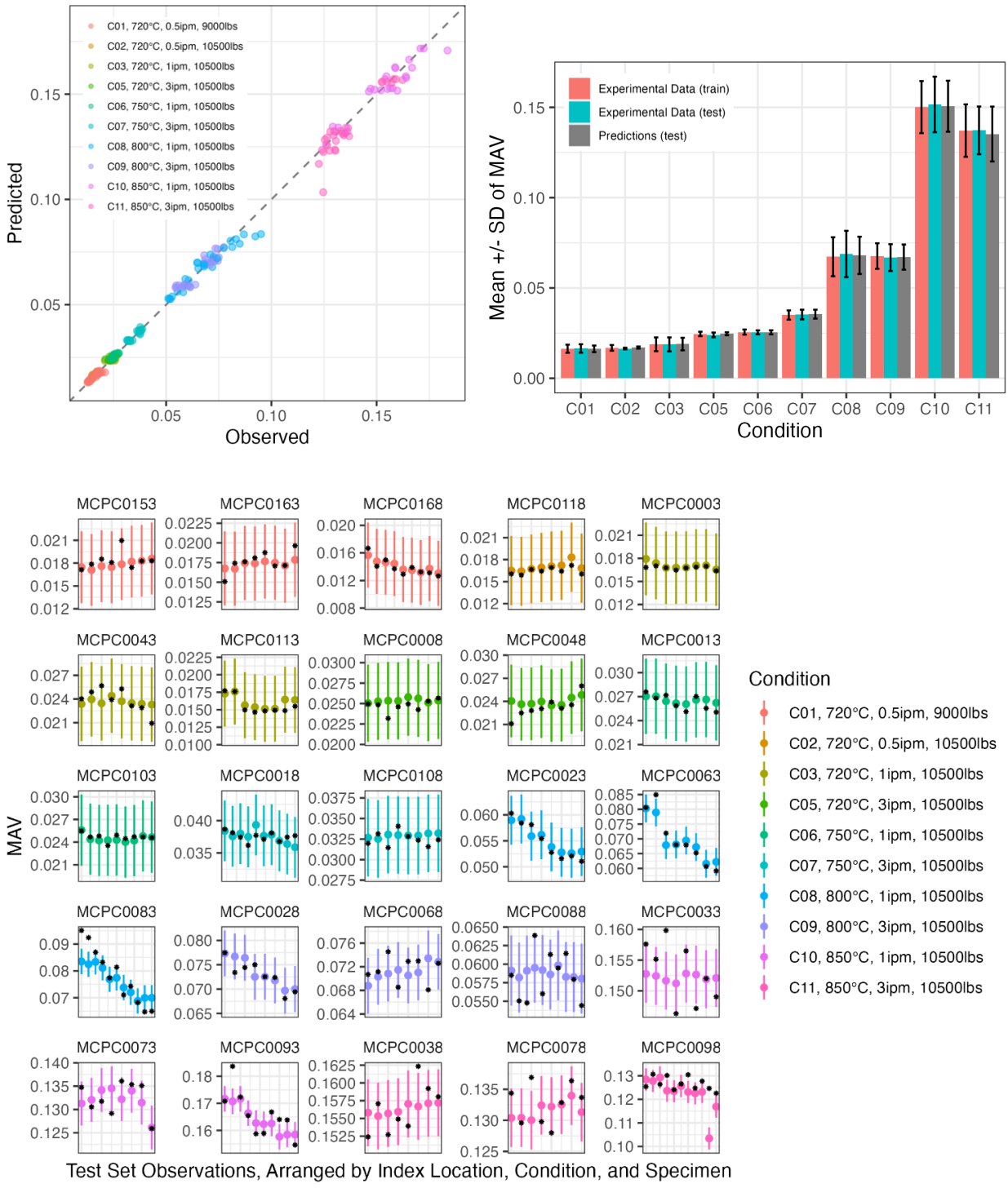


Figure 29. Predicted versus observed plots reveal overall good predictive skill of the RF model (top-left); mean +/- standard deviation of experimental and predicted data (top-right); and test set observations and predictions with PIs (bottom). Black stars are observed MAV.

We used the fitted RF model to extend predictions to grain size. It is of interest to investigate the effect of different definitions of MAV. Because the grain size is at the specimen level, not index

level (corresponding to a specific traverse position), we theoretically considered MAV to be the average of the raw voltages within the specimen, across indices within both the training and testing sets. In practice, we computed MAV as the mean of the index-level MAV values from the training sets, as the test set is assumed to be unobserved. Figure 30 shows a log-log model fitting grain size to MAV defined in both senses. Examining the coefficients establishes that the different definitions of MAV (theoretical and practical) have little effect on the relationship as the intercept and coefficient are barely changed. Standard 95% PIs for the log-log model are plotted below.

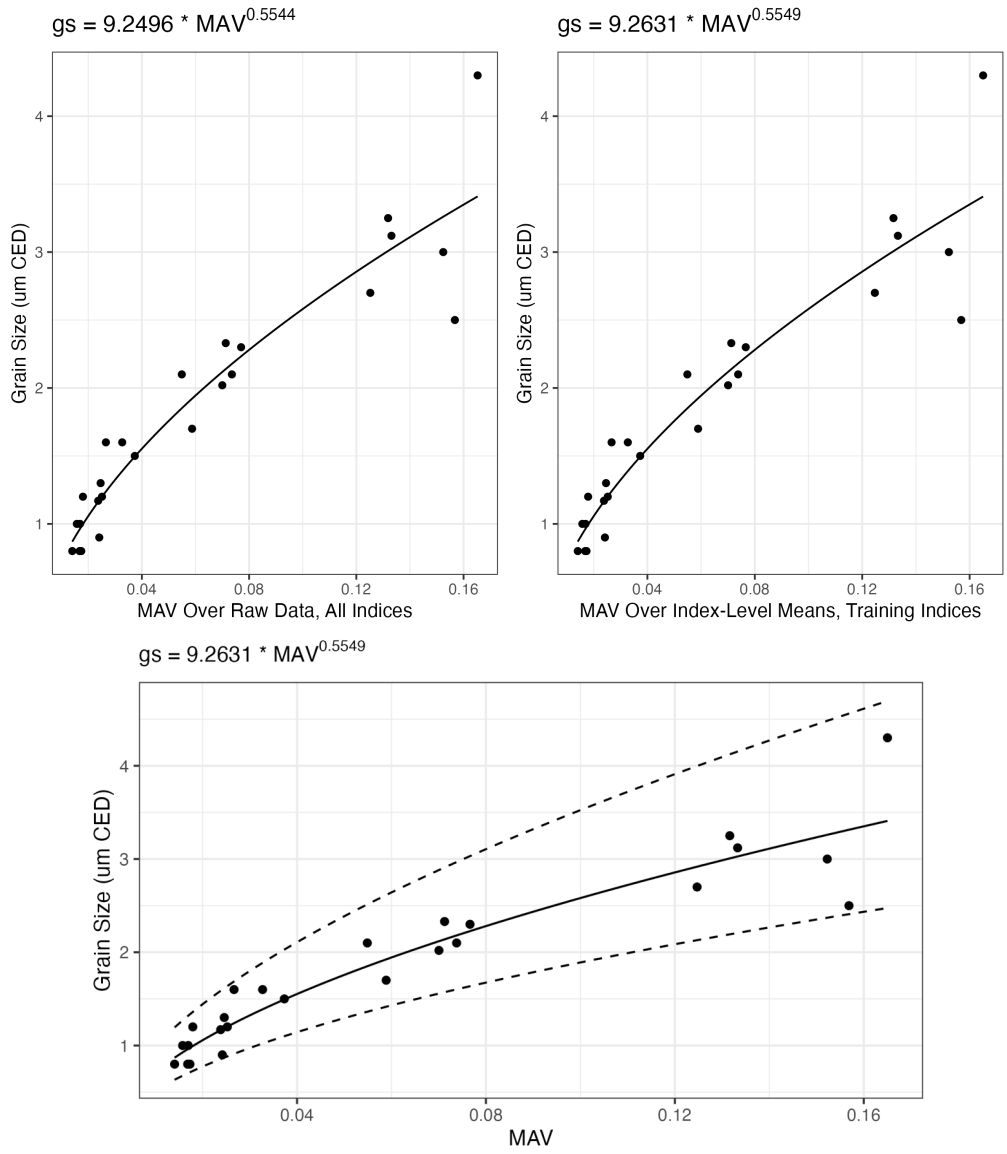


Figure 30. Log-log model fitted on the aggregated raw data (top-left) and the average of index-level averages (top-right); and log-log model with 95% PIs where the MAV is computed over index-level means for training indices only.

The PIs from the MAV predictions were mapped to the log-log PIs to obtain a rough propagation of error, demonstrated in Figure 31. 95% PIs for the RF MAV predictions and 95% PIs for the log-log model allow for an overall 90% confidence level according to the Bonferroni correction.

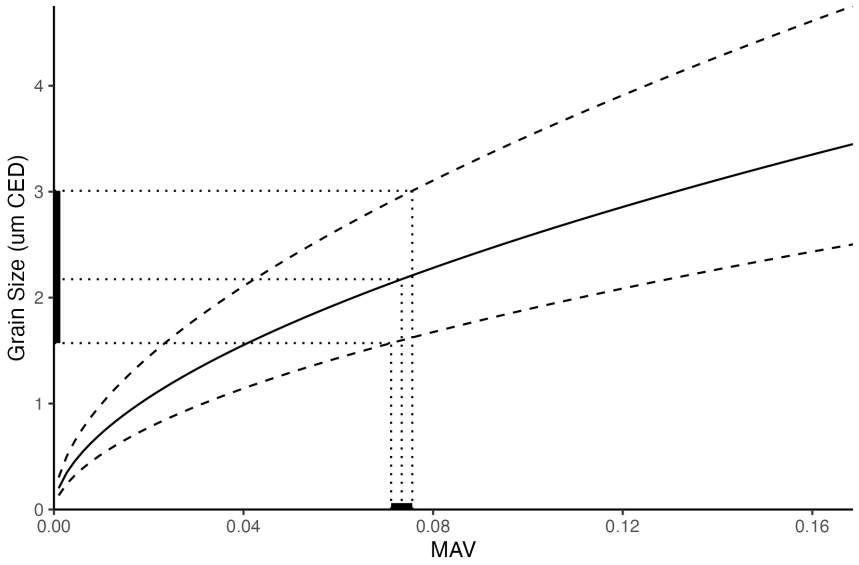


Figure 31. Log-log model with overall 90% PIs, mapping predicted MAV for MCPC Specimen 28 to predicted grain size.

Our goal in this task was to predict MAV (UT transducer signals) as a function of process parameters. We choose to focus on MAV at the index level to consider simple constructions and aggregations to support ease of interpretation before extension to more complex approaches. We included all MCPC specimens in both the train and test sets, as the goal was to predict MAV and obtain inferential information about process parameters across all conditions and index locations. However, in some cases, such as anomaly detection or statistical process control, it may be preferred to hold out entire specimens or regions of specimens.

We learned after initial analysis that the local FSP process parameters were highly correlated. Thus, we considered methods for highly correlated predictors, such as Elastic Net. However, no model is fully robust against highly correlated predictors, and an optimal solution would be to initially select a sensible subset of variables based on expert input. RF and lasso/ridge (the components of Elastic Net) regression can deal with high-dimensional data but can suffer in contexts with highly correlated predictors. Lasso can randomly drop important variables, and for RF, interpreting important variables can be opaque (Gregorutti et al. 2017). Thus, without additional steps to deal with correlation (e.g., dropping highly correlated variables based on systems/materials knowledge, using a recursive variable selection, employing ensemble models), our results are more appropriate for prediction rather than interpretation.

Because the number of observations within each condition is relatively small, all data was used for a single model across all conditions. When constructing PIs, this means that the errors for all conditions are of the same width for all conditions. This results in over-conservative PIs for the low conditions and under-conservative intervals for the higher conditions. Test observations at lower conditions (C01–C07) are covered by the PIs at a rate of 100%, while the observations at higher conditions are more likely to fall outside the PIs. The selection of a Bonferroni correction for prediction confidence is highly conservative (VanderWeele and Mathur 2019). The PIs could be modified by (1) collecting more data to allow for fitting separate models per condition and (2) using a different confidence correction (e.g., Bonferroni-Holm (Holm 1979)).

The extension of MAV predictions to grain size is possible, as shown in this task, but undesirable as there are now two steps, both introducing error into the final conclusion. If the final goal is grain size prediction, we should use grain size as the model response directly to minimize error. While in our case there are very few grain size observations (one per specimen for a total of 29 scalar observations), it would be possible with more data to predict grain size as a function of specimen-level process parameters. In this work, the PIs are presented without proof, but future work could explore theoretical results for the bootstrapped PIs. Other models could be considered that address the need for propagating errors in independent variables, such as errors-in-variables models (Stoklosa et al. 2023).

5.0 Conclusions and Future Work

This section provides a global conclusion for the work performed in the MCPC data analytics Vertex and highlights the future work that may be pursued. We note that each individual section in sections 3.0 and 4.0 provides section-specific conclusions associated with the related tasking.

In the DA Vertex of the MCPC Agile, we developed a framework consisting of data-driven approaches to associate global and/or local process parameters, microstructural features, and properties of FSP 316L SS plates. The models developed in this effort describe patterns, trends, and correlations in the associated datasets within the bounds of the training data. Two separate frameworks were developed in this Vertex, namely the PREDICT and CONTROL modules. The PREDICT framework was developed to accelerate materials and process pathways discovery—it can identify the microhardness and grain size distribution of FSP 316L SS plates processed under specific global conditions. Alternatively, the PREDICT framework can also identify the global process parameters, such as the combination of FSP tool temperature and traverse rate that could result in desirable grain size distribution in the processed plate or the local microhardness. The CONTROL framework was developed to consistently achieve desirable part qualities during SPP. It can predict the local process conditions, such as required tool rotation rate, forge force, and power that must be maintained to ensure the desired local average grain size in an FSP 316L SS plate.

The dataset used to develop these frameworks was complex, comprising data as tables, graphs, images, and categorical labels. As such, to develop the desired correlations between the various independent and dependent variable datasets, different models were employed. Therefore, the PREDICT and CONTROL modules consist of prediction frameworks that are composites of multiple data-driven models based on classic statistical analysis, ML and DL, with significant data flow between them as required. Only experimental data was used to assemble these frameworks in the current effort. However, the frameworks can be extended to include computationally derived datasets in the future as well.

Specifically, the objective of the PREDICT framework was to develop PSP relations for FSP 316L SS plates using image and numerical data. The PREDICT framework consists of four different models that could associate PSP datasets.

1. We developed a generative AI model conditioned on microstructural images from the nugget region of the FSP 316L SS plates as a function of the global FSP parameters, namely tool traverse rate and tool temperature. We used normalizing flows, a generative AI model, to generate synthetic images of FSP 316L microstructures after training the model on real local SEM images of the processed nugget region of the FSP 316L SS plates. For comparison, we also developed a GAN to generate synthetic images of the FSP 316L SS plate conditioned on the same process conditions. We noted that normalizing flows outperformed GANs in terms of feature integrity. Notably, the normalizing flow model showed ~80% accuracy in generating images conditioned on FSP forge force and >55% accuracy in generating temperature-conditioned microstructures. We also showed in this task that training on a larger number of images with lower magnification and therefore a lower number of details does not result in high model performance. We showed that training on a smaller dataset of images with higher magnification (requiring longer data acquisition times) results in better predictions.
2. We developed a microstructure segmentation tool based on UNet++ architecture. We developed correlated SEM–EBSD images for training the models, which consist of labels for the grains and grain boundary regions in the microstructural images of the nugget region of

the 316L plates. We identified a high density of subgranular boundaries in the SPP 316L, which is typically not seen in annealed microstructures. This required identification of grain boundaries extensively in the EBSD software. We trained the segmentation models using a topological loss function on datasets consisting of backscatter electron and secondary electron images collected at different accelerating voltages. Our results showed that most of the model architectures we developed predicted the ECD of an image in the range predicted by EBSD. We also noted that training on secondary electron images at a lower accelerating voltage provided the lowest MAE in ECD predictions. We showed that the UNet++ architecture could achieve errors in predicting ECD as low as 0.5 μm and an accuracy of ~70%.

3. We developed a series of statistical, DL, and ML models to predict the numerical average ECD of SS grains of a processed FSP microstructure as a function of global FSP parameters, namely tool traverse rate, tool temperature (collected by a thermocouple at the tool shoulder), forge force applied during processing, and the tool location along the processing direction. Our results showed that the statistical LMM showed an accuracy of >80%, which is like that achieved by the symbolic regression DL model as well. Both the statistical and DL models could also define the mathematical relationships between the process parameters and the grain size distribution. However, the PI was tighter for the symbolic regression model compared to the LMM. The ML model, RF, used to predict the ECD also showed >80% accuracy and was able to identify tool temperature as the most significant factor among the process conditions in predicting the grain size.
4. Like numerical ECD prediction, to predict the microhardness of the processed region of the FSP 316L SS plate at a specific location, we developed similar statistical, DL, and ML models. The LMM statistical model, the symbolic regression DL model, and the RF ML model all showed >80% accuracy in predicting the average microhardness at a location in the processed region of the FSP 316L plate as a function of tool traverse rate, tool temperature (collected by a thermocouple at the tool shoulder), forge force applied during processing, and the tool location along the processing direction. Both the statistical and DL models could also define the mathematical relationships between the process parameters and the grain size distribution. However, the PI was tighter for the symbolic regression model compared to the LMM. Here, specifically, we explored the effects of providing location-specific properties vs. global properties to the ML model. Our results showed that training the RF model with the locations at which microhardness was collected increased model accuracy significantly to >95%.

   Whereas the PREDICT framework focused on global PSP relationships, the CONTROL framework developed was aimed at predicting specific local process parameters as a function of other available data during FSP or local microstructural features such as grain size as a function of NDE data. As such, the CONTROL framework consists of three different models, described below.

5. We developed a physics informed neural network (PINN) to predict the local tool temperature as a function of the powder required during processing that is calculated as a function of the local torque and tool rotation rate. The PINN was developed as a universal differential equation paradigm with a regression task used to fit a dynamical system to data. The PINN was used to solve the lumped capacitance equation for a model dynamical system with input and output physics present in FSP, namely internal heat generation and heat transfer away from the processed region. Seven FSP experiments were recorded (i.e., having full temperature time histories) with consistent material selections and similar dwell times. Given this limited amount of data, and with a goal of training an interpretable and

generalizable model, we employ a data augmentation strategy and a low-data model validation strategy. We noted that the lumped-capacitance abstraction was sufficient to capture essential physics for reproducing behavior seen in the input power data as a function of time (time series data). Because the model was constructed with constituent energy pathways, the model is interpretable in that the mapping from tool force and temperature to heat input is available.

6. We developed LSTM models to predict the local processed region depth and tool rotation rate seen during FSP in the steady-state processing region as a function of the local process parameters experienced previously. The modeling paradigms we examined were a single-step history model that used an MLP, a time-stepper LSTM model that outputs the actual value, and a residual time-stepper LSTM model that outputs the rate of change from the initial condition consisting of the time series process parameter at a location and instant. We explored the effects of hidden dimensions and time steps on prediction accuracy and concluded that with the training data. We found that the residual model performed worse than the exact value time-stepper model, and the models with memory for the entire sequence had higher performance with an average error in prediction of tool rotation rate as low as 4 rpm.
7. We developed an RFM associating local FSP parameters with mean average voltage obtained from UT NDE performed on sections of FSP 316L SS plates under steady-state conditions. The model showed a low mean average percentage error in prediction of ~4%. We learned after initial analysis that the local FSP process parameters were highly correlated and considered methods for highly correlated predictors understanding that RF's identification of important variables is rather opaque. Thus, the results are more appropriate for prediction rather than interpretation. We then developed a statistical correlation between the UT NDE data and the ECD of the processed region obtained from EBSD imaging of the samples, resulting in a hybrid ML-statistical model. The extension of MAV predictions to grain size was possible but undesirable as there are now two steps, both introducing error into the conclusion. Despite this, the log-log relationship developed between MAV and the ECD could capture the associations between the dataset within a 95% PI.

The PREDICT and CONTROL frameworks are the first comprehensive step in developing AI tools for accelerating research and development of new materials, components, and demonstrations or controlling manufacturing processes using datasets created specifically for training and developing the models. Previous frameworks developed on PNNL's Mathematics and Artificial Reasoning Sciences and Solid Phase Processing Science initiative investments used legacy manufacturing and material data that was previously collected for different purposes. While the PREDICT and CONTROL frameworks did just that—provide the framework of AI tools and information flow—they will need to be further developed and updated to increase the accuracy of predictions in order to deploy them successfully. Industrial manufacturing and shop floor environments require >99% accuracy in property prediction or process envelope forecasting in order to use the models in real-time quality assurance/quality control activities. There are two approaches currently available for achieving enhanced accuracy of ML models. One is to optimize model parameters in the latent space and how the data is provided to the model during development to optimize for minimizing model error (MAE, or other similar factors). However, the most effective way to achieve higher model accuracy in many cases is to enhance the quality of the data available for training. We note here that not all instances require big data. Previous efforts showed small datasets can yield high-accuracy predictions if the datasets are curated effectively, while focusing on model tuning for low errors in predictions. Some of the tasking in the DA Vertex also explored the effects of training ML and DL models on a large dataset with lower quality vs. a smaller dataset with higher quality of details and found

that the second option yielded better accuracy in model predictions aligning with experimentally obtained ground truth.

Developing a high-quality dataset is a very challenging endeavor, as we discovered during the MCPC Agile. However, it is extremely important to achieve high-performance AI models, as we previously discussed. The challenge is in the definition of what makes a good dataset and from whose perspective. Experimentalists have traditionally stewarded data development (if the dataset consists of experimentally developed data) or computational experts (if the dataset consists of simulated data). As we discovered during the model developed in the DA Vertex in the MCPC Agile, a good dataset developed by either of these parties may not be considered a sufficiently complete or good dataset by data scientists. This is especially true in research and development environments where material and process development are sometimes viewed (and rightfully so) as an art rather than a science since many new directions explored are based on an intuitive estimation of the next thing to do rather than an exhaustive exploration of every factor available in optimizing an outcome. The discrepancy typically lies in two different areas. The first one is the extent of the details that need to be collected for a single experiment or simulation—while the data generators may disregard a few data streams as superfluous, the data scientist's perspective may be to ask for an exhaustive data collection approach. This may also be seen as the quality of the data collected from an experiment or a simulation with a focus on details on all streams of data—which is usually a resource-intensive undertaking but becomes necessary all the same. The second one is an understanding of the number of classes of data and the extent that it should be made available for model development—what we term as data spread. For developmental activities, data is typically collected through random sampling initially, followed by focused data collection in certain conditions, which prompt toward achieving the desired outcomes. However, in datasets required for AI model training, it might be better to (a) develop data points in the extremes of the process envelopes and (b) when possible, use uniform sampling to collect the data in the rest of the process envelopes. These steps may assure a dataset that is more balanced, which is supportive of realistic predictions from the AI models.

One drastic change required from the status quo, then, is that data collection should be a collaborative effort between data generators and the data users (who eventually build the AI models). Instead of using only legacy data that was collected independently of the data users or regarding data collection as a one-shot activity pursued at the beginning of a project, we recommend making data collection a collaborative endeavor between the data generators and data users. We encouraged such a collaboration in the MCPC Agile, where data scientists were a part of designing the data generation activities. During this collaboration, we learned a dataset designed to have good, bad, and ugly data points was much better at achieving high accuracy predictions from the models built than legacy datasets that have an overrepresentation of good outcomes. We also learned that standardization of data collection is key for developing a good dataset that the data scientists in model building can curate and use easily. Finally, we learned and recommend for future activities that data collection has to be seen as an iterative process during the course of model development where additional information should be provided as needed to build accurate, dependable models. All of this required a hand-in-hand partnership between data generators and data consumers.

Another key finding from the DA Vertex was the need for continued collaboration between data users and generators through the course of model development. It is never sufficient to only know what the independent and dependent variables are to build an AI model. Every iteration of model development requires the involvement of data generators to be involved to (1) provide understanding of what the data columns mean physically and how to use them during model

development, and (2) if the results output from the AI models make sense. We note here that data-driven models do not have physics principles built into them, as expected. Therefore, models provide results, which in certain cases, may prove to not be physically viable. For example, during the MCPC Agile, we noted that one DL model predicted negative tool rotation rates as the result of correlating it with local process parameters—which is not achievable physically. Similarly, another ML model predicted that tool temperatures >1400 °C were ideal for achieving average ECD of ~5 µm based on the training data provided—which is not possible since 316L SS melts at 1375 °C. In these instances, experimentalists reviewing the model outcomes were quick to point out the issues with the model prediction, which led to the development of physics-constrained models in the DA Vertex, where artificial limitations were placed on predictions to make sure the outcomes were physically viable and reasonable, even before they were confirmed experimentally. This was possible because the data generators worked closely with the AI model developers to interpret the data physically during model building as well as understand the implications of model outcomes during development, which led to varying model development and optimization strategies.

Discovering materials with desired properties (such as hardness, fatigue strength, creep strength, corrosion resistance, or ultimate tensile strength) or microstructural features (such as grain size, crystallographic orientation, degree of consolidation (% porosity), or precipitate density and morphology) is a great capability in the larger context of materials science and manufacturing research. Recent developments suggest AI tools, such as those developed in the DA Vertex of the MCPC Agile, will be crucial in accelerating their discovery and manufacturing process development in the larger context. In the future, we anticipate extending the model development to be based not only on experimentally derived data but also on computationally developed data streams. For example, smooth particle hydrodynamics is a great approach for representing the complex material evolution during SPP but is computationally resource- and time intensive. Developing data-driven, AI-based lower-order models that are physics constrained and can capture the associations in the independent and dependent variables would be a great direction to pursue. Such models can not only provide the insights of smooth particle hydrodynamics models at lower computing loads and higher speeds but also be used to fill in the data gaps left behind by experimental data collection. We also identify that the CONTROL framework developed in the DA Vertex only provided the basis for predicting local process parameters required to maintain properties or microstructural features within the desired gamut in a material or component being manufactured. What is necessary in the future is to develop control programs based on the CONTROL framework process parameters and demonstrate the effectiveness of AI-based model predictive control programs in achieving desired component quality. We note that the models developed in this effort were the first demonstrations for developing associations and predicting microstructural features and properties. We anticipate they will need to be optimized further and possibly combined to provide multiple outcomes from a single model rather than have the frameworks consist of composite predictions that feed into each other. We anticipate the prediction errors to decrease with the use of more complex DL models that can forecast multiple outcomes using the same input datasets that are expandable to accommodate newer datasets—which requires a robust approach to data management as well. Finally, we anticipate the PREDICT and CONTROL frameworks to be adopted in the future by users who may not be data scientists—they may find use in developing the design of experiments for research and development endeavors by material scientists or be employed by equipment operators to program control programs during manufacturing. Thus, it will be crucial for all the models developed currently and in the future on the current dataset to embed interpretability or explainability parameters to build trust with the model users for their continued utilization.

6.0 References

Bachmann, F., R. Hielscher, and H. Schaeben. 2011. “Grain detection from 2d and 3d EBSD data—Specification of the MTEX algorithm.” *Ultramicroscopy*, 111(12):1720-1733. DOI: https://doi.org/10.1016/j.ultramic.2011.08.002.

Bahari-Sambran, F., F. Carreño, C. M. Cepeda-Jiménez, and A. Orozco-Caballero. 2024. “Predicting grain size-dependent superplastic properties in friction stir processed ZK30 magnesium alloy with machine learning methods.” *Journal of Magnesium and Alloys,* 12(5):1931-1943. https://doi.org/10.1016/j.jma.2024.05.019.

Bland, J. M. and D. G. Altman. 1995. “Multiple significance tests: The Bonferroni method.” *British Medical Journal*, 310(6973):170.

Boldsaikhan, E., E. M. Corwin, A. M. Logar, and W. J. Arbegast. 2011. “The use of neural network and discrete Fourier transform for real-time evaluation of friction stir welding.” *Applied Soft Computing*, 11(8):4839–4846. doi: 10.1016/j.asoc.2011.06.017.

Breiman, L. 2001. “Random Forests.” *Machine Learning*, 45(1):5-32. https://doi.org/10.1023/A:1010933404324.

Chang, C. I., C. J. Lee, and J. C. Huang. 2004. “Relationship between grain size and Zener–Holloman parameter during friction stir processing in AZ31 Mg alloys.” *Scripta materialia* 51(6):509-514. https://doi.org/10.1016/j.scriptamat.2004.05.043.

Chen, T. and C. Guestrin. 2016. “XGBoost: A scalable tree boosting system.” In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. https://doi.org/10.1145/2939672.2939785.

Chen, R., Y. Rubanova, J. Bettencourt, and D. Duvenaud. 2018. “Neural Ordinary Differential Equations.” *Advances in Neural Information Processing Systems,* 31.

Chowdhury, S. A., M. F. N. Taufique, J. Wang, M. Masden, M. Wenzlick, R. Devanathan, A. L. Schemer-Kohrn, and K. S. Kappagantula. 2024. “Automated Grain Boundary (GB) Segmentation and Microstructural Analysis in 347H Stainless Steel Using Deep Learning and Multimodal Microscopy.” *Integrating Materials and Manufacturing Innovation*, 13(1):244-256. doi: 10.1007/s40192-023-00305-7.

Ede, J. M. 2021. “Deep learning in electron microscopy.” *Machine Learning: Science and Technology*, 2(1):011004. doi: 10.1088/2632-2153/abd614.

Garcia, D., T. Wang, J.D. Escobar, M. Pole, X. Ma, and K. A. Ross. 2024. “In-situ measurement and control of the tool-workpiece interface temperature during friction stir processing of 304/304L stainless steel.” *Materials Today Communications*, 3:107672. doi: https://doi.org/10.1016/j.mtcomm.2023.107672.

Goldstein, J. I., D. E. Newbury, J. R. Michael, N. W. Ritchie, J. H. J. Scott, and D. C. Joy. 2017. *Scanning Electron Microscopy and X-ray Microanalysis*. Springer, Springer New York, NY. https://doi.org/10.1007/978-1-4939-6676-9.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. 2014. “Generative adversarial nets.” *Advances in Neural Information Processing Systems*, 27. (NIPS 2014) Curran Associates, Inc.

Gregorutti, B., B. Michel, and P. Saint-Pierre. 2017. “Correlation and variable importance in random forests.” *Statistics and Computing*, 27(3):659-678. https://doi.org/10.1007/s11222-016-9646-1.

Haitovsky, Y. 1969. “Multicollinearity in regression analysis: Comment.” *The Review of Economics and Statistics*, 51(4):486-489. https://doi.org/10.2307/1926450

Hartl, R., A. Bachmann, J. B. Habedank, T. Semm, and M. F. Zaeh. 2021. “Process monitoring in friction stir welding using convolutional neural networks.” *Metals*, 11(4):535. https://doi.org/10.3390/met11040535.

Hastie, T., R. Tibshirani, and M. Wainwright. 2015. *Statistical learning with sparsity: The Lasso and generalizations*. Chapman and Hall/CRC. https://doi.org/10.1201/b18401

Heidarzadeh, A., M. Jabbari, and M. Esmaily. 2015. “Prediction of grain size and mechanical properties in friction stir welded pure copper joints using a thermal model.” *The International Journal of Advanced Manufacturing*, 77, 1819–1829 (2015). https://doi.org/10.1007/s00170-014-6543-7

Heidarzadeh, A. .S. Mironov, R. Kaibyshev, G. Çam, A. Simar, A. Gerlich, F. Khodabakhshi, A. Mostafaei, D. P. Field, J. D. Robson, et al. 2021. “Friction stir welding/processing of metals and alloys: A comprehensive review on microstructural evolution.” *Progress in Materials Science*, 117:100752. doi: https://doi.org/10.1016/j.pmatsci.2020.100752.

Hill, R. C., and G. Judge. 1987. “Improved prediction in the presence of multicollinearity.” *Journal of Econometrics*, 35(1):83-100.

Hirabayashi, Y., H. Iga, H. Ogawa, S. Tokuta, Y. Shimada, and A. Yamamoto. 2024. “Deep learning for three-dimensional segmentation of electron microscopy images of complex ceramic materials.” *npj Computational Materials*, 10(1):46. doi: 10.1038/s41524-024-01226-5.

Hochreiter, S. and J. Schmidhuber. 1997. “Long Short-Term memory.” *Neural Computation*, 9(8):1735-1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

Holm, S. 1979. “A simple sequentially rejective multiple test procedure.” *Scandinavian Journal of Statistics*, 6(2):65–70.

Hu, X. F. Li, D. Samaras, and C. Chen. 2019. “Topology-preserving deep image segmentation. Advances in neural information processing systems.” In *Proceedings of the Advances in Neural Information Processing Systems 32 (NEURIPS 2019)*. https://papers.nips.cc/paper_files/paper/2019/hash/2d95666e2649fcfc6e3af75e09f5adb9-Abstract.html.

Humphreys, F. J. 2001. “Review Grain and subgrain characterisation by electron backscatter diffraction.” *Journal of Materials Science*, 36(16):3833-3854. doi: 10.1023/A:1017973432592.

Hynes, N. R. J., P. Nagaraj, and J. A. J. Sujana. 2014. “Ultrasonic evaluation of friction stud welded AA 6063/AISI 1030 steel joints.” *Materials & Design* (1980-2015), 62:118–123. https://doi.org/10.1016/j.matdes.2014.05.017

Jazaeri, H. and F. J. Humphreys. 2004. “Quantifying recrystallization by electron backscatter diffraction.” *Journal of Microscopy*, 213(3):241-246. https://doi.org/10.1111/j.0022-2720.2004.01296.x.

Jia, J. M., M. Staring, and B. C. Stoel. 2022. “seg-metrics: a Python package to compute segmentation metrics.” *medRxiv* 2024, doi: https://doi.org/10.1101/2024.02.22.24303215.

Kingma, D. P. and J. L. Ba. 2014. “Adam: A Method for Stochastic Optimization.” *arXiv* (Cornell University), Jan. 2014, doi: 10.48550/arxiv.1412.6980.

Koch, J., W. Choi, E. King, D. Garcia, H. Das, T. Wang, K. Ross, and K. Kappagantula. 2024. “Neural lumped parameter differential equations with application in friction-stir processing.” *Journal of Intelligent Manufacturing,* 36:1111-1121. doi:10.1007/s10845-023-02271-5

Kuhn, M. and H. Wickham. 2020. Tidymodels: A collection of packages for modeling and machine learning using tidyverse principles. [Computer software]. https://www.tidymodels.org.

Kuhn, M. and K. Johnson. 2013. *Applied Predictive Modeling*. Springer. New York, NY https://doi.org/10.1007/978-1-4614-6849-3.

Kuhn, M. and K. Johnson. 2020. 3. “A Review of the Predictive Modeling Process.” In *Feature Engineering and Selection: A Practical Approach for Predictive Models*. CRC Press Taylor and Francis Group.

Lloyd, G. E. 1987. “Atomic number and crystallographic contrast images with the SEM: a review of backscattered electron techniques.” *Mineralogical Magazine*. 51(359):3-19. doi:10.1180/minmag.1987.051.359.02.

Legaard, C. T. Schranz, G. Schweiger, and J. Drgona. 2023. “Constructing neural network based models for simulating dynamical systems.” *ACM Computing Surveys*, 55(11):1-34. doi: 10.1145/3567591.

Liu, F. C. and T. W. Nelson. 2017. “In-situ grain structure and texture evolution during friction stir welding of austenite stainless steel.” *Materials & Design*, 115:467-478. doi: https://doi.org/10.1016/j.matdes.2016.11.066.

Liu, F. C. and T. W. Nelson. 2016. “In-situ material flow pattern around probe during friction stir welding of austenitic stainless steel.” *Materials & Design* 110:354-364. doi: https://doi.org/10.1016/j.matdes.2016.07.147.

Liu, T. and J. Regier. 2020. “An Empirical Comparison of GANs and Normalizing Flows for Density Estimation.” *arXiv* preprint arXiv:2006.10175 (2020). https://doi.org/10.48550/arXiv.2006.10175.

Mironov, S., Y. S. Sato, H. Kokawa, H. Inoue, and S. Tsuge. 2011. “Structural response of superaustenitic stainless steel to friction stir welding.” *Acta Materialia*, 59(14):5472-5481. doi: https://doi.org/10.1016/j.actamat.2011.05.021.

Mishra, R. S. P. S. De, and N. Kumar. 2014. “Fundamentals of the Friction Stir Process.” In *Friction Stir Welding and Processing: Science and Engineering*, Springer International Publishing, 2014; pp 13-58.

Mishra, R. S. and Z. Y. Ma. 2005. “Friction stir welding and processing.” *Materials Science and Engineering: R: Reports*, 50(1):1-78. doi: https://doi.org/10.1016/j.mser.2005.07.001.

Morgan, D., G. Pilania, A. Couet, B. P. Uberuaga, C. Sun, and J. Li. 2022. “Machine learning in nuclear materials research.” *Current Opinion in Solid State and Materials Science*. 26(2):100975. doi: https://doi.org/10.1016/j.cossms.2021.100975.

Muhamedyev, R., K. Yakunin, Y. A. Kuchin, A. Symagulov, T. Buldybayev, S. Murzakhmetov, and A. Abdurazakov. 2020. “The use of machine learning ‘black boxes’ explanation systems to improve the quality of school education.” *Cogent Engineering*, 7(1). doi: 10.1080/23311916.2020.1769349.

Nowell, M. M., R. A. Witt, and B. W. True. 2005. “EBSD Sample Preparation: Techniques, Tips, and Tricks.” *Microscopy Today*, 13(4):44-49. https://doi.org/10.1017/S1551929500053669.

Patala, S., J. K. Mason, and C. A. Schuh. 2012. “Improved representations of misorientation information for grain boundary science and engineering.” *Progress in Materials Science*, 57(8):1383-1425. doi: https://doi.org/10.1016/j.pmatsci.2012.04.002.

Petch, J., S. Di, and W. Nelson. 2022. “Opening the Black Box: The Promise and Limitations of Explainable Machine Learning in Cardiology.” *Canadian Journal of Cardiology*, 38(2). doi: 10.1016/j.cjca.2021.09.004.

PNNL. 2024. Solid Phase Processing Demonstration Facility.” Pacific Northwest National Laboratory, Richland, Washington. https://www.pnnl.gov/solid-phase-processing-demonstration-facility.

Pyzer-Knapp, E. O., J. W. Pitera, J. W. J. Staar, S. Takeda, T. Laino, D. P. Sanders, J. Sexton, J. R. Smith, and A. Curioni. 2022. “Accelerating materials discovery using artificial intelligence, high performance computing and robotics.” *npj Computer Materials* 8:84. https://doi.org/10.1038/s41524-022-00765-z.

Rajamanickam, N., V. Balusamy, G. Madhusudhanna Reddy, and K. Natarajan, “Effect of process parameters on thermal history and mechanical properties of friction stir welds.” *Materials & Design*, 30(7): 2726-2731. doi: 10.1016/j.matdes.2008.09.035.

Rudin, C. 2019. “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.” *Nature Machine Intelligence*, 1(5):206-215. doi: 10.1038/s42256-019-0048-x.

Saxena, P., A. Bongale, and S. Kumar. 2022. “A review of friction stir processing in fabricating surface composites for process control and sensor data monitoring.” *Materials Today: Proceedings*, 50:2099-2104. https://doi.org/10.1016/j.matpr.2021.09.427.

Shen, C., C. Wang, M. Huang, N. Xu, S. van der Zwaag, and W. Xu. 2021. “A generic high-throughput microstructure classification and quantification method for regular SEM images of

complex steel microstructures combining EBSD labeling and deep learning." *Journal of Materials Science & Technology*, 93:191-204. doi: https://doi.org/10.1016/j.jmst.2021.04.009.

Smith, E. L. 2021. LDRD Project 79593 (AGILE) Project Management Plan Lite.

Stoklosa, J., W.-H. Hwang, and D. I. Warton. 2023. "A general algorithm for error-in-variables regression modelling using Monte Carlo expectation maximization." *PLOS ONE*, 18(4). e0283798. https://doi.org/10.1371/journal.pone.0283798.

Stuckner, J., B. Harder, and T. M. Smith. 2022. "Microstructure segmentation with deep learning encoders pre-trained on a large microscopy dataset." *npj Computational Materials*, 8(1):200. doi: 10.1038/s41524-022-00878-5.

Sudhagar, S., M. Sakthivel, and P. Ganeshkumar. 2019. "Monitoring of friction stir welding based on vision system coupled with Machine learning algorithm." *Measurement*, 144:135-143, doi: 10.1016/j.measurement.2019.05.018.

Sun, H., P. Ramuhalli, and R. E. Jacob. 2023. "Machine learning for ultrasonic nondestructive examination of welding defects: A systematic review." *Ultrasonics*, 127, 106854. https://doi.org/10.1016/j.ultras.2022.106854

Tabatabaeipour, M., J. Hettler, S. Delrue, and K. Van Den Abeele. 2016. "Non-destructive ultrasonic examination of root defects in friction stir welded butt-joints." *NDT & E International*, 80:23-34. https://doi.org/10.1016/j.ndteint.2016.02.007.

Todd, D. R., K. Nwe, M. Pole, Y. Guo, A. D. Guzman, A. Ortiz, and M. J. Olszta, et al. 2024a. *Materials Characterization, Prediction, and Control Project; Summary Report on Material Characterization, Part 1*. PNNL-36768. Pacific Northwest National Laboratory, Richland, WA.

Todd, D. R., K. Nwe, M. Pole, Y. Guo, A. D. Guzman, A. Ortiz, and M. J. Olszta, et al. 2024b. *Materials Characterization, Prediction, and Control Project; Summary Report on Material Characterization, Part 2*. PNNL-36963. Pacific Northwest National Laboratory, Richland, WA.

Todd D. R., K. Nwe, M. Pole, Y. Guo, A. D. Guzman, A. Ortiz, and M. J. Olszta, et al. 2024c. *Materials Characterization, Prediction, and Control Project; Summary Report on Material Characterization, Part 3*. PNNL-36964. Pacific Northwest National Laboratory Richland, WA.

Todd D. R., K. Nwe, M. Pole, Y. Guo, A.D. Guzman, A. Ortiz, and M. J. Olszta, et al. 2024d. *Materials Characterization, Prediction, and Control Project; Summary Report on Material Characterization, Part 4*. PNNL-36965. Pacific Northwest National Laboratory Richland, WA.

Tong, V. S. and T. B. Britton. 2021. "TrueEBSD: Correcting spatial distortions in electron backscatter diffraction maps." *Ultramicroscopy*, 221:113130. doi: https://doi.org/10.1016/j.ultramic.2020.113130.

VanderWeele, T. J. and M. B. Mathur. 2019. "Some desirable properties of the Bonferroni correction: Is the Bonferroni correction really so bad?" *American Journal of Epidemiology*, 188(3):617–618. https://doi.org/10.1093/aje/kwy250.

Verma, S., V. Msomi, S. Mabuwa, A. Merdji, J. Prakash Misra, U. Batra, and S. Sharma. 2021. "Machine learning application for evaluating the friction stir processing behavior of dissimilar

aluminium alloys joint." In *Proceedings of the Institution of Mechanical Engineers Part L Journal of Materials Design and Applications*, 236(3):146442072110531. doi: 10.1177/14644207211053123.

Wang, B., S. Zhong, T.-L. Lee, K. S. Fancey, and J. Mi. 2020. "Non-destructive testing and evaluation of composite materials/structures: A state-of-the-art review." *Advances in Mechanical Engineering*, 12(4), 1687814020913761. https://doi.org/10.1177/1687814020913761.

Wang, D., D. R. Ni, B. L. Xiao, Z. Y. Ma, W. Wang, and K. Yang. 2014. "Microstructural evolution and mechanical properties of friction stir welded joint of Fe–Cr–Mn–Mo–N austenite stainless steel." *Materials & Design* 64:355-359. doi: https://doi.org/10.1016/j.matdes.2014.07.063.

Wang, T., D. Garcia, M. Pole, and K. A. Ross. 2024."Force reduction of friction stir welding and processing of steel." *Materialia*, 33, 102050. doi: https://doi.org/10.1016/j.mtla.2024.102050.

Wickham, H. 2019. "Welcome to the {tidyverse}." *Journal of Open Source Software*, 4(43):1686. https://doi.org/10.21105/joss.01686.

Wight, C. L., S. M. Akers, W. Choi, T. H. Emerson, L. J. Gosink, E. R. Jurrus, and K. S. Kappagantula, et al. 2022. "DeepTemp: Predicting Material Processing Conditions with Artificial Intelligence." Presented by C.L. Wight at Artificial Intelligence in Materials and Manufacturing 2022, Pittsburgh, Pennsylvania. PNNL-SA-171530.

Zhang, H., J. Zimmerman, D. Nettleton, and D. J. Nordman. 2020. "Random Forest prediction intervals." *The American Statistician*. 74(4):392-406. https://www.tandfonline.com/doi/abs/10.1080/00031305.2019.1585288

Zhou, Z., M. R. Siddiquee, N. Tajbakhsh, and J. Liang. 2018. "UNet++: A Nested U-Net Architecture for Medical Image Segmentation." Cham, 2018; Springer International Publishing: pp 3-11. https://doi.org/10.1007/978-3-030-00889-5.

7.0 Appendix 1 – Friction stir process parameters

The base material for this work was 316L stainless steel plate. The plate dimensions were 152.4 × 330.2 × 15.9 mm (6 × 36 × 0.625 in.). All FSP was performed parallel to the rolling direction. The mill scale and surface oxide were removed from the surface using a hand-held die grinder with 180 grit abrasive sandpaper. The friction stir tool was a convex scrolled shoulder stepped spiral (CS4) tool purchased from Mazak MegaStir with an E44230A design. The tool had a shoulder diameter of 23.4 mm, pin height of 5.7 mm height, and pin root diameter of 10.7 mm. The tool was made from polycrystalline cubic boron nitride (PCBN) with 30 wt.% W-Re. FSP for this work was performed on a custom Manufacturing Technology, Inc. friction stir weld machine at Pacific Northwest National Laboratory. A simultaneous and independent temperature and force control algorithm was used during each run. As such, the independent variables for this work were the processing temperature, the traverse velocity, and the forging force. The tool rotation rate and plunge depth were dependent variables that were controlled by the algorithm. In brief, a constant temperature was obtained by maintaining a power setpoint through direct control of the tool rotation rate. Note that the temperature was measured by a thermocouple positioned at the back of the PCBN, so this is not representative of the tool-workpiece interface temperature. The processing temperatures selected for this work were, 720, 750, 800, and 850 °C. Similarly, a forging force control algorithm was used to maintain a setpoint for all FSP runs in this work. Three traverse velocities were used: 0.21 (0.5 ipm), 0.42 mm/s (1 ipm) and 1.27 mm/s (3 ipm). The tilt angle for this work was 0.5°. A quartz shield with continuous argon flow surrounded the FSP tool to prevent oxidation during processing. The quartz shield had a diameter of 95.25 mm (3.75 in.) diameter. The processing conditions for all experiments are summarized in Table 2. In a single 316L stainless steel plate (with a specific ID), FSP was performed under four different conditions as shown in the figure below. Multiple samples were manufactured with the same processing parameters held for a condition to assess repeatability. Subsamples were subsequently cut from the sample manufactured under a specific processing condition and provided for hardness testing, microscopy, and ultrasonic testing. Table A1 provides the IDs for the plates and samples per process condition for reference.

Table A1. Nomenclature for 316L plates on which FSP samples were manufactured per conditions listed.

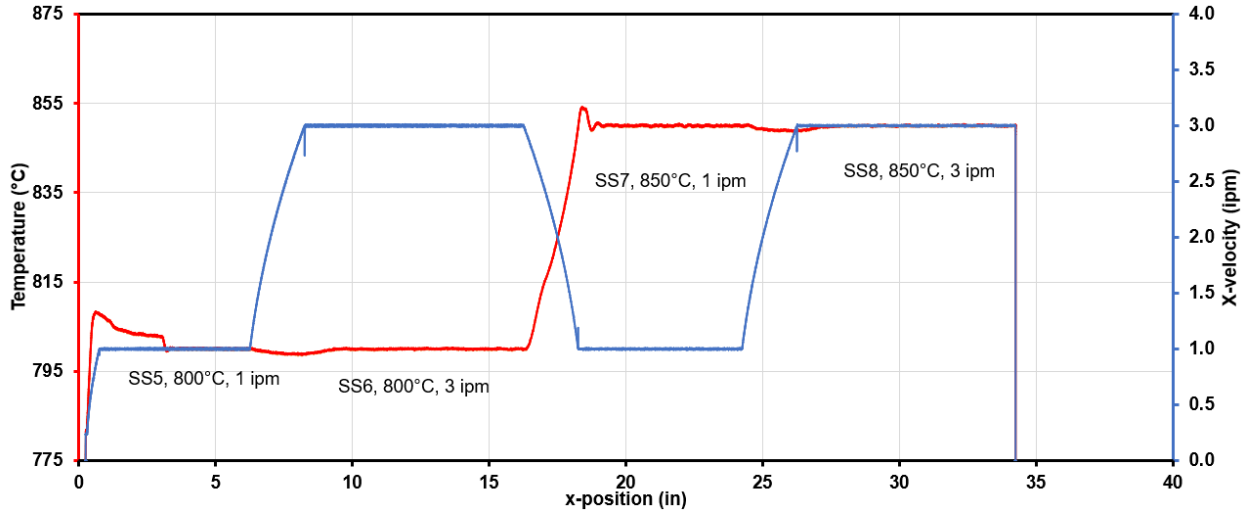
Condition	Sample ID	Temperature	Traverse velocity (in/min)	Force limit (lb)	Plate ID
C00	SS36	N/A	N/A	N/A	N/A
C01	SS31	720	0.5	9000	2022 03 07 0013
	SS33	720	0.5	9000	2022 03 08 0003
	SS34	720	0.5	9000	2022 03 08 0003
	SS35	720	0.5	9000	2022 03 08 0003
C02	SS28	720	0.5	10500	2022 03 07 0011

	SS24	720	0.5	10500	2022 03 07 0009
C03	SS1	720	1	10500	2022 03 02 05
	SS9	720	1	10500	2022 03 02 0017
	SS23	720	1	10500	2022 03 07 0009
	SS27	720	1	10500	2022 03 07 0011
C04	SS32	720	1	9000	2022 03 07 0013
C05	SS2	720	3	10500	2022 03 02 05
	SS10	720	3	10500	2022 03 02 0017
C06	SS3	750	1	10500	2022 03 02 05
	SS11	750	1	10500	2022 03 02 0017
	SS21	750	1	10500	2022 03 07 0009
	SS25	750	1	10500	2022 03 07 0011
	SS29	750	1	10500	2022 03 07 0013
C07	SS4	750	3	10500	2022 03 02 05
	SS12	750	3	10500	2022 03 02 0017
	SS22	750	3	10500	2022 03 07 0009
	SS26	750	3	10500	2022 03 07 0011
	SS30	750	3	10500	2022 03 07 0013
C08	SS5	800	1	10500	2022 03 02 0009
	SS13	800	1	10500	2022 03 02 0019
	SS17	800	1	10500	2022 03 04 0005/0007
C09	SS6	800	3	10500	2022 03 02 0009
	SS14	800	3	10500	2022 03 02 0019
	SS18	800	3	10500	2022 03 04 0005/0007
C10	SS7	850	1	10500	2022 03 02 0009
	SS15	850	1	10500	2022 03 02 0019

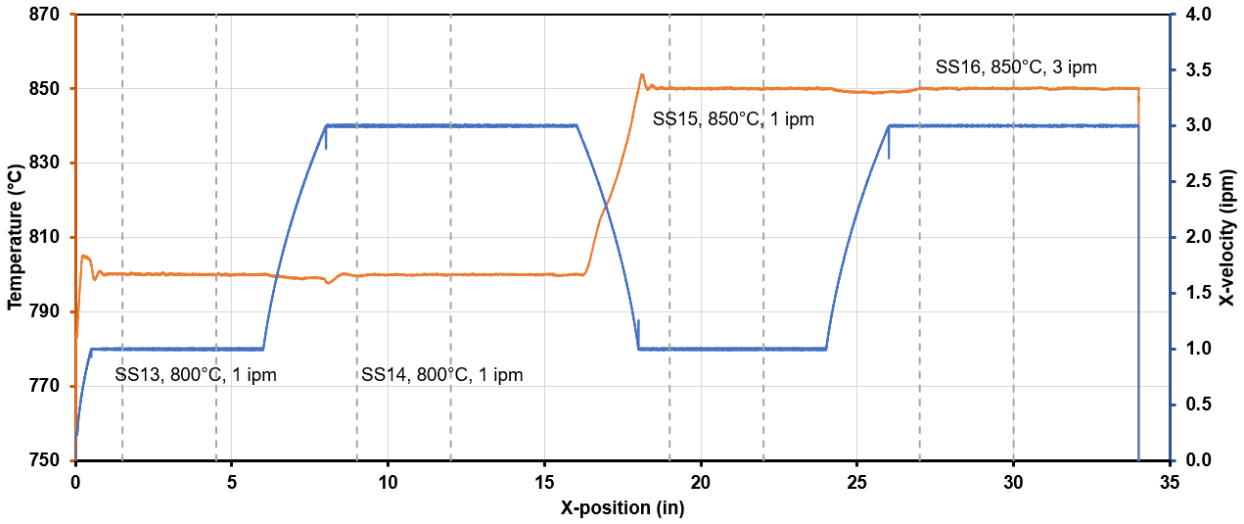
	SS19	850	1	10500	2022 03 04 0005/0007
C11	SS8	850	3	10500	2022 03 02 0009
	SS16	850	3	10500	2022 03 02 0019
	SS20	850	3	10500	2022 03 04 0005/0007

Figure A1 below shows the variation in tool temperature and traverse velocity (x-velocity) as a function of traverse position (x-position) along the length of the weld during FSP of the 316L plates.

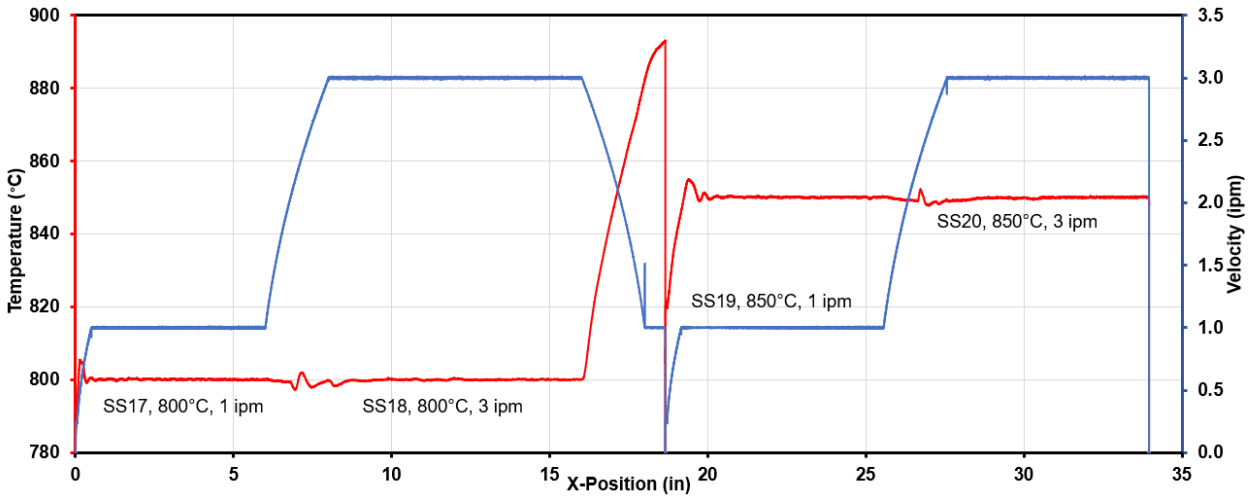
A. Plate – 2022 03 02 0009 (C08-C11)



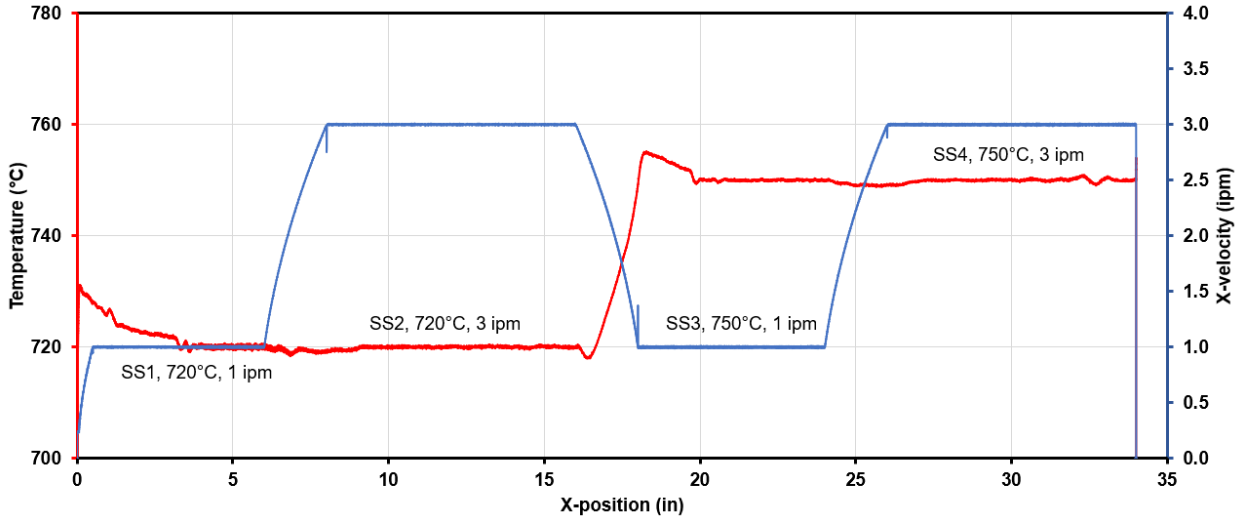
B. Plate – 2022 03 02 #19 (C08 – C11)



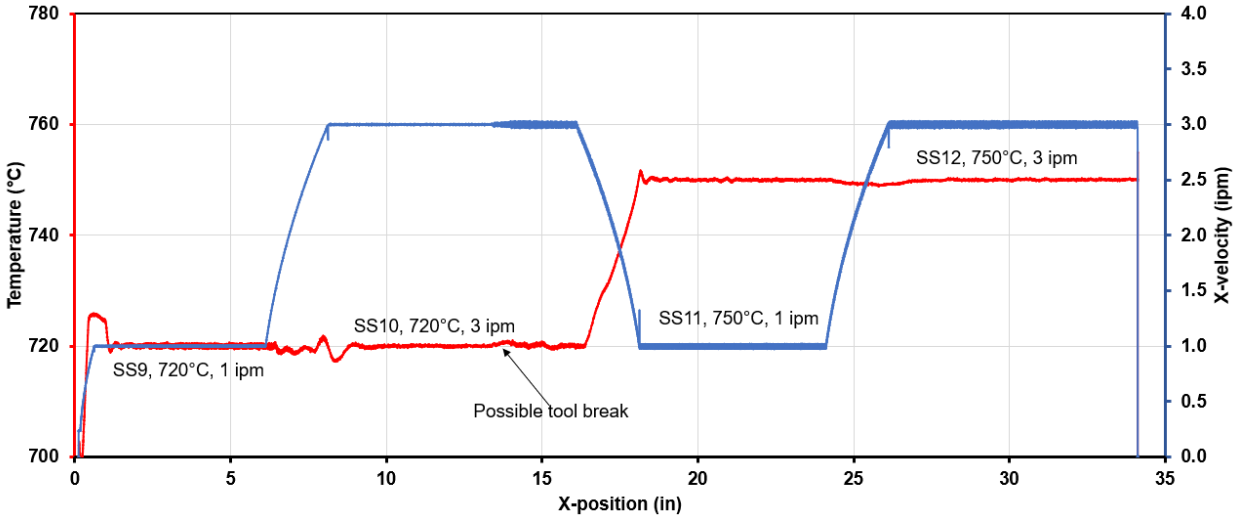
C. Plate – 2022 03 04 0005/0007 (C08 – C11)



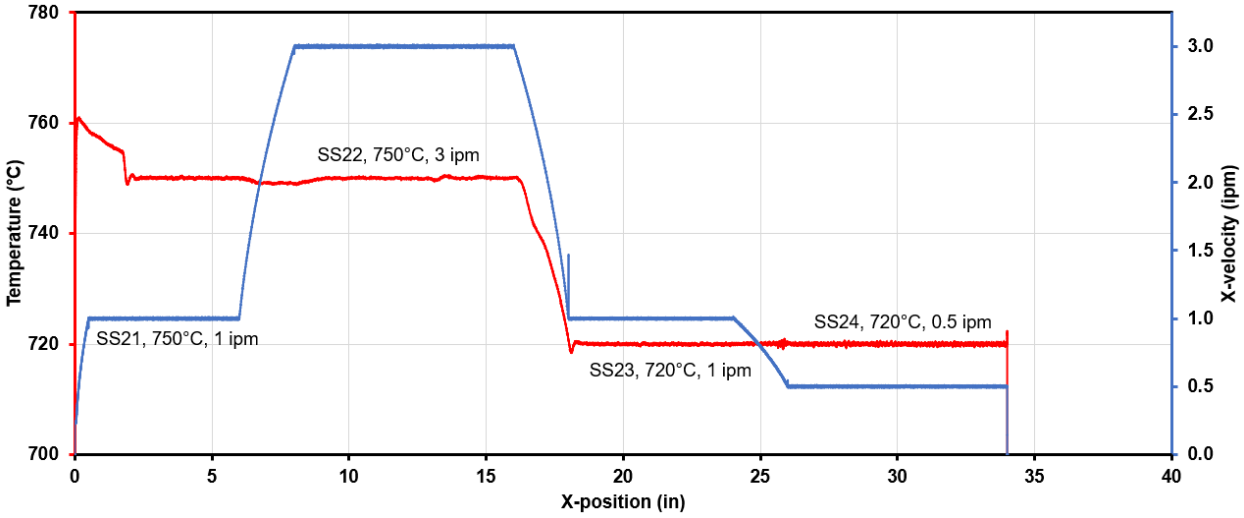
D. Plate – 2022 03 02 05 (C03, C05, C06, C07)



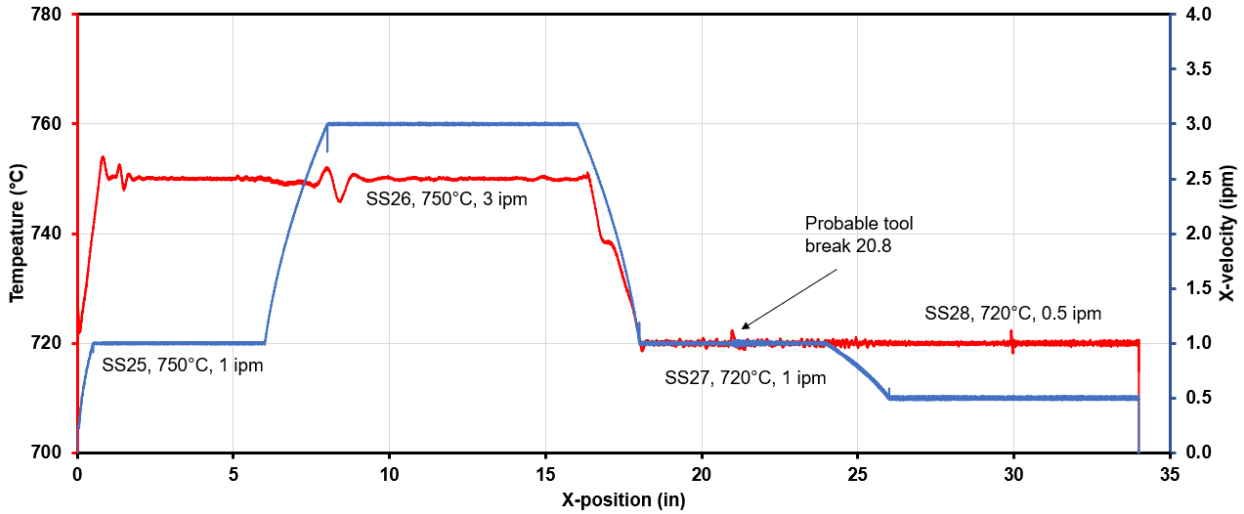
E. Plate – 2022 03 02 0017 (C03, C05, C06, C07), 720-750°C



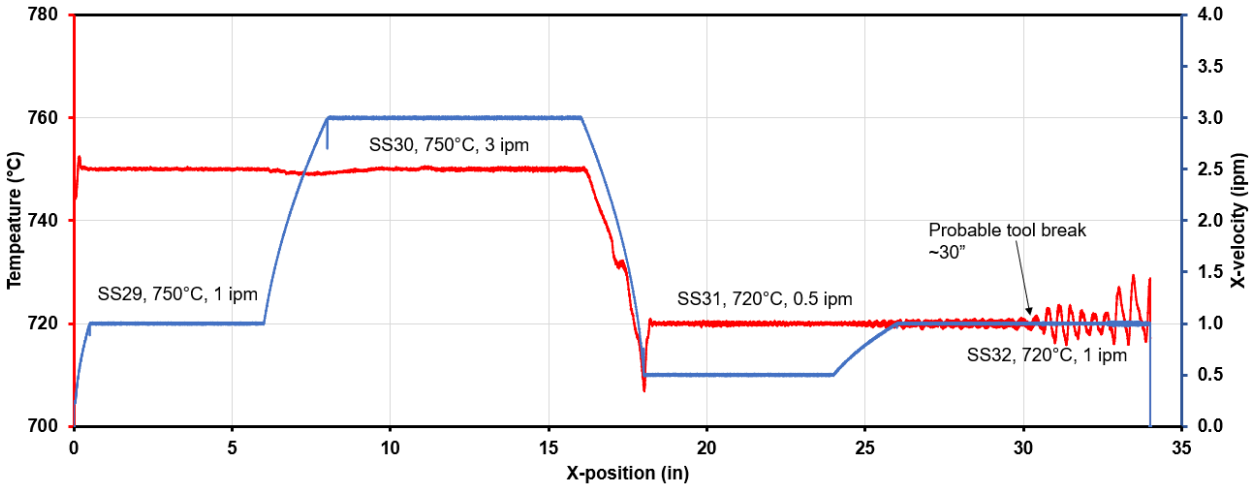
F. Plate – 2022 03 07 0009 (C06, C07, C03, C02)



G. Plate – 2022 03 07 0011 (C06, C07, C03, C02)



H. Plate – 2022 03 07 0013 (C06, C07, C01, C04)



I. Plate – 2022 03 08 0003 (C01) 720°C, 9.0klbs

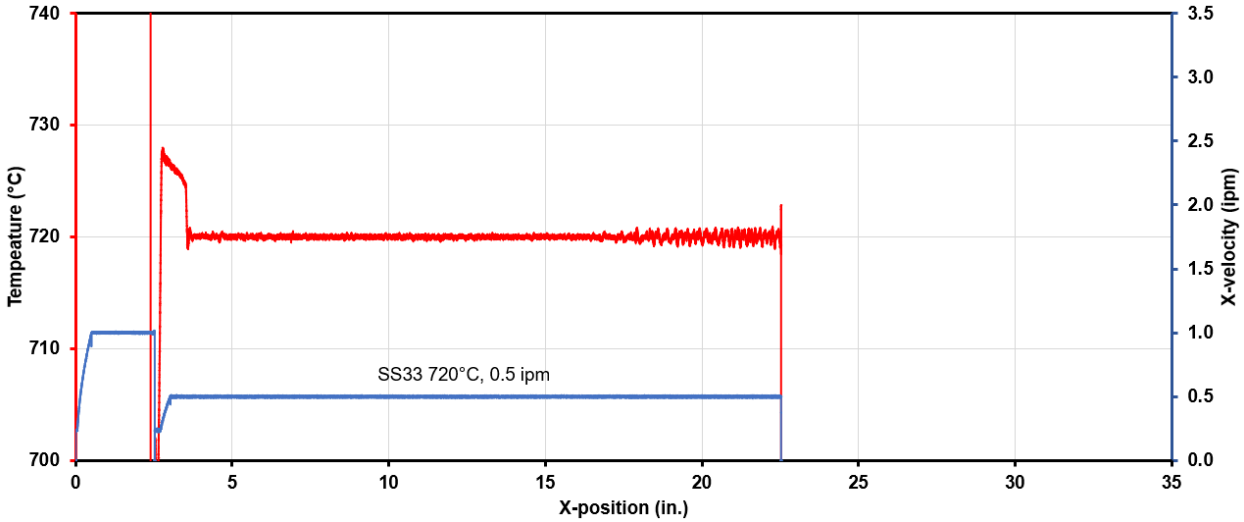


Figure A1. Variation in temperature and tool traverse velocity (x-velocity) as a function of traverse location (x-position) during FSP of 316L stainless steel plates.

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354

1-888-375-PNNL (7665)

www.pnnl.gov