Journal Name

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

Supplementary Information For: Exploring the Transferability of Machine-Learning Models for Analyzing XRD Data of Shocked Microstructures: From Single Crystal to Polycrystals

Daniel Vizoso,[a] Phillip Tsurkan,[b] Ke Ma,[b] Avinash M. Dongare,[b] and Rémi Dingreville[a‡]

S1 Machine-learning model architectures

S1.1 Convolutional Autoencoder Architecture and Training

The convolutional autoencoder used for this work is composed of an encoding model and a symmetric decoding model. The encoding model is composed of three convolutional layers, each with a kernel size of 5. The specific encoder architecture is defined as: $\mathrm{Conv}^{\mathrm{ReLU}}_{4\times121} \times \mathrm{Conv}^{\mathrm{ReLU}}_{8\times117} \times \mathrm{Conv}^{\mathrm{ReLU}}_{16\times113} \times \mathrm{Flat}_{1\times1808} \times \mathrm{Linear}_{1808\times L_D=20}$. The nomenclature “Conv" describes a convolutional layer, “Linear" a dense linear layer, and “Flat" a flattening operation. The subscript represents the dimension of a single input after being fed through each layer, with the first number indicating the number of channels and the second number indicating the number of components in each channel. The superscript indicates the activation function that was used after its respective layer, with the ReLU activation function being used after each convolutional operation in the encoder. The architecture of the decoding portion of the autoencoder is a mirrored version of the architecture of the encoder, with the exception that the ReLU activation function is called after the initial inverse linear operation, and is not called after the final transpose convolutional operation. The network architecture has been selected based on prior work [1,2] which demonstrated that this general architecture was appropriate to perform a regression task on similar problems, using similar data formats.

Training of the autoencoder was performed using the mean squared error (MSE) function as the loss function, where the value of the loss was computed by calculating the mean squared difference between the input XRD profiles and the reconstructed XRD profiles produced by the decoder. Optimization of the autoencoder's parameters was performed using the AdamW optimizer as implemented in PyTorch, with a learning rate of 10^{-3} and a weight decay of 10^{-5}, with parameter optimization being performed over a total of 1 000 training epochs.

S1.2 MLP Architecture and Training

The MLP is composed of six linear layers that gradually expand the number of dimension up until the fourth layer, followed by two linear layers that contract the number of dimensions down to the number of microstructural descriptors that the model is trained to predict. The specific architecture of the MLP is defined as: $\mathrm{Linear}^{\mathrm{ReLU}}_{1\times32} \times \mathrm{Linear}^{\mathrm{ReLU}}_{1\times32} \times \mathrm{Linear}^{\mathrm{ReLU}}_{1\times64} \times \mathrm{Linear}^{\mathrm{ReLU}}_{1\times128} \times \mathrm{Drop}_{p=0.25} \times \mathrm{Linear}^{\mathrm{ReLU}}_{1\times32} \times \mathrm{Linear}_{1\times6}$. The nomenclature “Drop" refers to a dropout layer, with the subscript indicating the probability p that determines what fraction of the data at that layer is set to zero. The dropout layer is used as a regularization technique to prevent overfitting during training. As in the architecture of the autoencoder, the ReLU activation function is used after each linear layer except for the last layer.

Training of the MLP was also performed using the MSE function as the loss function, with the loss being computed by comparing the known microstructural state descriptor values to the predicted values produced by the MLP. Optimization of the MLP's parameters was also performed using the AdamW optimizer, with a learning rate of 10^{-4} and a weight decay of 10^{-5}. The number of training epochs for the MLP was adjusted based on the composition of the training dataset, as certain shock-loading orientations or combinations of orientations were more prone to overfitting than others. When trained with the complete set of XRD profiles captured from the $\langle 111 \rangle$ loading direction single-crystal simulation which contained 51 different timesteps, the MLP was trained for 1 000 epochs. For all other single-crystal loading direction shock simulations as well as for a truncated set from the $\langle 111 \rangle$ loading direction simulation that only included the same timesteps as the other loading directions, the MLP was trained for 3 000 epochs.

[a] *Center for Integrated Nanotechnologies, Sandia National Laboratories, NM, USA*
[b] *Department of Materials Science and Engineering, Institute of Materials Science, University of Connecticut, Storrs, CT, USA*
‡ email: rdingre@sandia.gov

S1.3 Training Datasets and Test-Train Splits

As listed in Table S1, a total of 11 different datasets were used for model training, with each dataset being composed of a single shock-loading simulation or multiple shock-loading simulations that used different loading orientations. For each training dataset, 11 different test-train splits were created in order to quantify how the creation of the test-train splits modified model training and prediction performance. The distribution of data between the test and training sets was kept the same when training the autoencoders and the MLPs. All models that included the ⟨111⟩ loading orientation as part of a set of multiple loading orientations used the truncated set (dataset 2), so that the number of timesteps included for each loading orientation would be the same. Test-train splits of 70% training and 30% test were used for model training, with a fixed batch size of 32 samples per batch.

Table S1 List of Datasets Used For Model Training

Set Number	Dataset Used For Model Training
1	⟨111⟩ (complete)
2	⟨111⟩ (truncated)
3	⟨110⟩
4	⟨100⟩
5	⟨112⟩
6	⟨111⟩, ⟨110⟩, ⟨100⟩, ⟨112⟩
7	⟨111⟩, ⟨110⟩, ⟨100⟩
8	⟨112⟩, ⟨110⟩, ⟨100⟩
9	⟨111⟩, ⟨110⟩, ⟨112⟩
10	⟨111⟩, ⟨100⟩, ⟨112⟩
11	Polycrystalline

S2 Microstructural state descriptor ranges and histograms for the binned regions of the shock-loading simulations

Table S2 Maximum and minimum pressures, total dislocation densities, phase fractions, and temperatures from the binned regions of the five different shock simulations

Microstructural Descriptor	Single Crystal ⟨111⟩	Single Crystal ⟨110⟩	Single Crystal ⟨100⟩	Single Crystal ⟨112⟩	Polycrystalline
Pressure (GPa)	-8.366 to 43.866	-6.864 to 52.750	-8.818 to 49.678	-9.565 to 50.314	-9.491 to 50.713
Total Dislocation Density ($Å^{-2}$)	0.0 to 12.304	0.0 to 13.066	0.0 to 3.819	0.0 to 8.665	0.326 to 6.083
Disordered Fraction	0.0 to 0.9495	0.0 to 0.7992	0.0 to 0.9683	0.0 to 0.8022	0.0847 to 0.6893
FCC Fraction	0.0381 to 1.0	0.1609 to 1.0	0.0195 to 1.0	0.1354 to 1.0	0.2310 to 0.9137
HCP Fraction	0.0 to 0.1514	0.0 to 0.1776	0.0 to 0.3510	0.0 to 0.2093	0.0010 to 0.30074
Temperature (K)	298.704 to 1425.779	299.419 to 1415.516	299.340 to 1622.331	299.425 to 1295.620	297.905 to 1369.995

Table S2 provides the maximum and minimum values for the six microstructure descriptors measured within the binned regions of the five different shock simulations performed for this work. As discussed in Section 3.1 of the main text, loading orientation and the initial microstructure had significant impacts on the evolution of the system during the shock-loading process for all of the measured microstructure state descriptors. Histograms of the distributions measured in the binned regions for the six microstructure state descriptors are provided in Figs. S1, S2, S3, S4, S5, and S6.

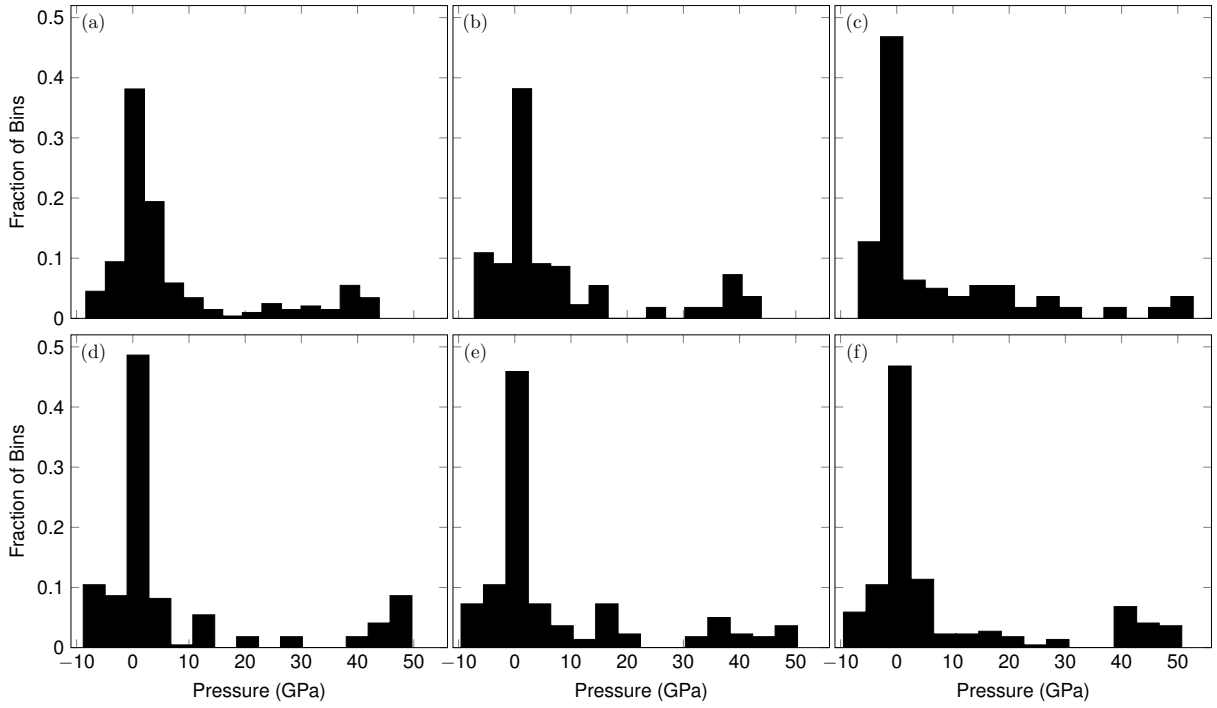

Fig. S1 Histograms of the pressures measured in the binned regions for set numbers: a) 1, b) 2, c) 3, d) 4, e) 5, f) 11.

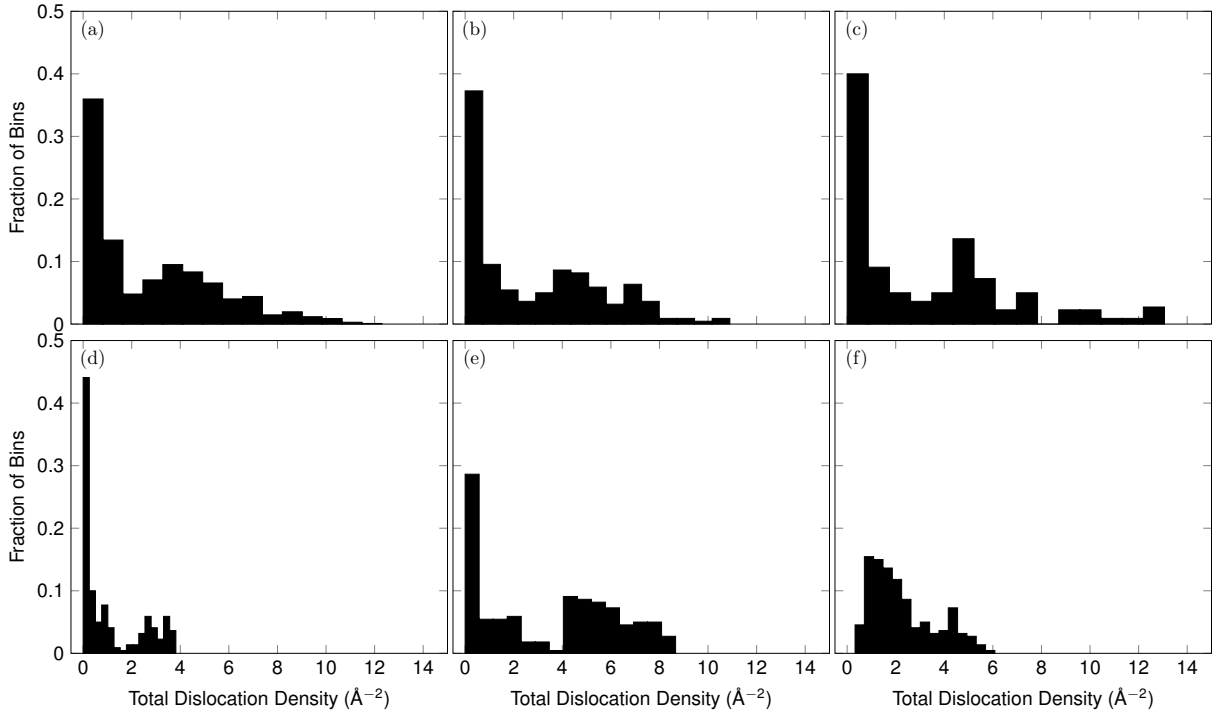

Fig. S2 Histograms of the total dislocation densities measured in the binned regions for set numbers: a) 1, b) 2, c) 3, d) 4, e) 5, f) 11.

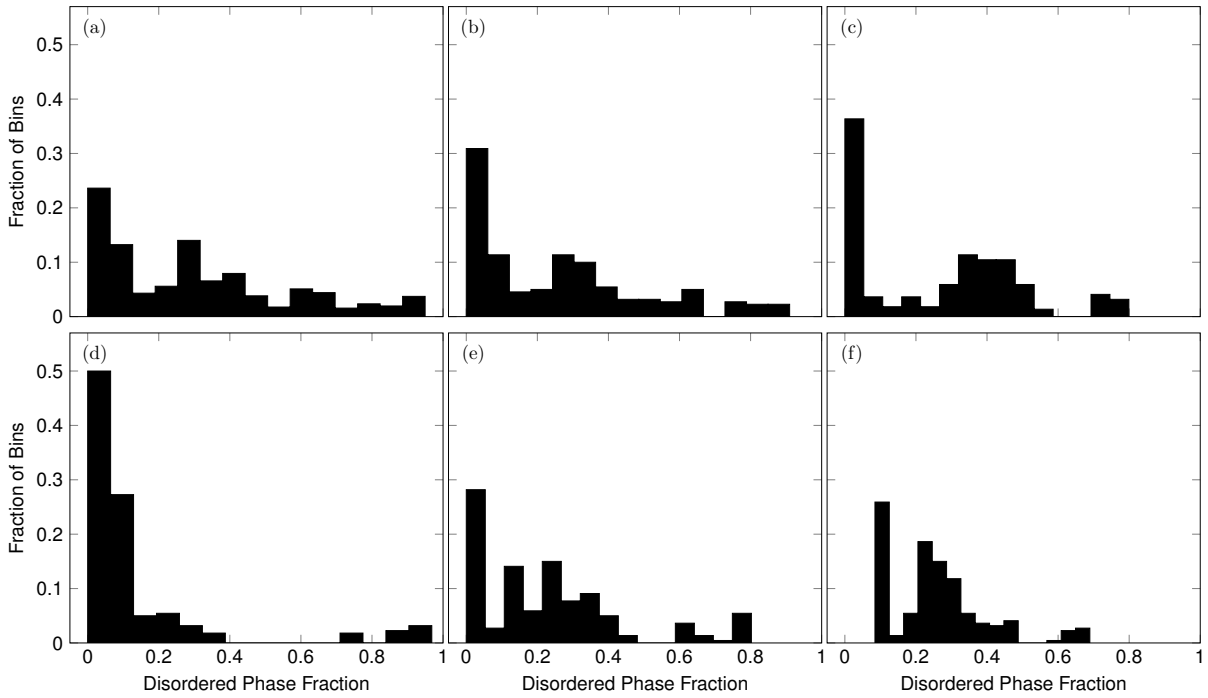

Fig. S3 Histograms of the disordered phase fractions measured in the binned regions for set numbers: a) 1, b) 2, c) 3, d) 4, e) 5, f) 11.

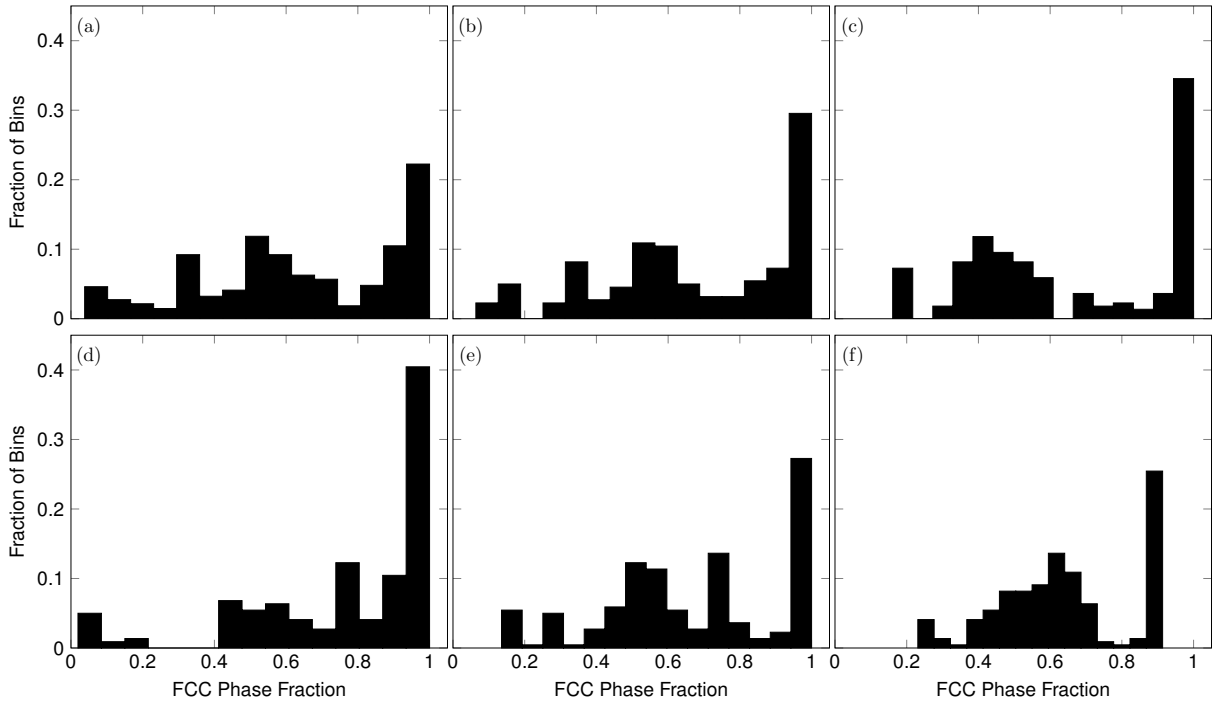

Fig. S4 Histograms of the FCC phase fractions measured in the binned regions for set numbers: a) 1, b) 2, c) 3, d) 4, e) 5, f) 11.

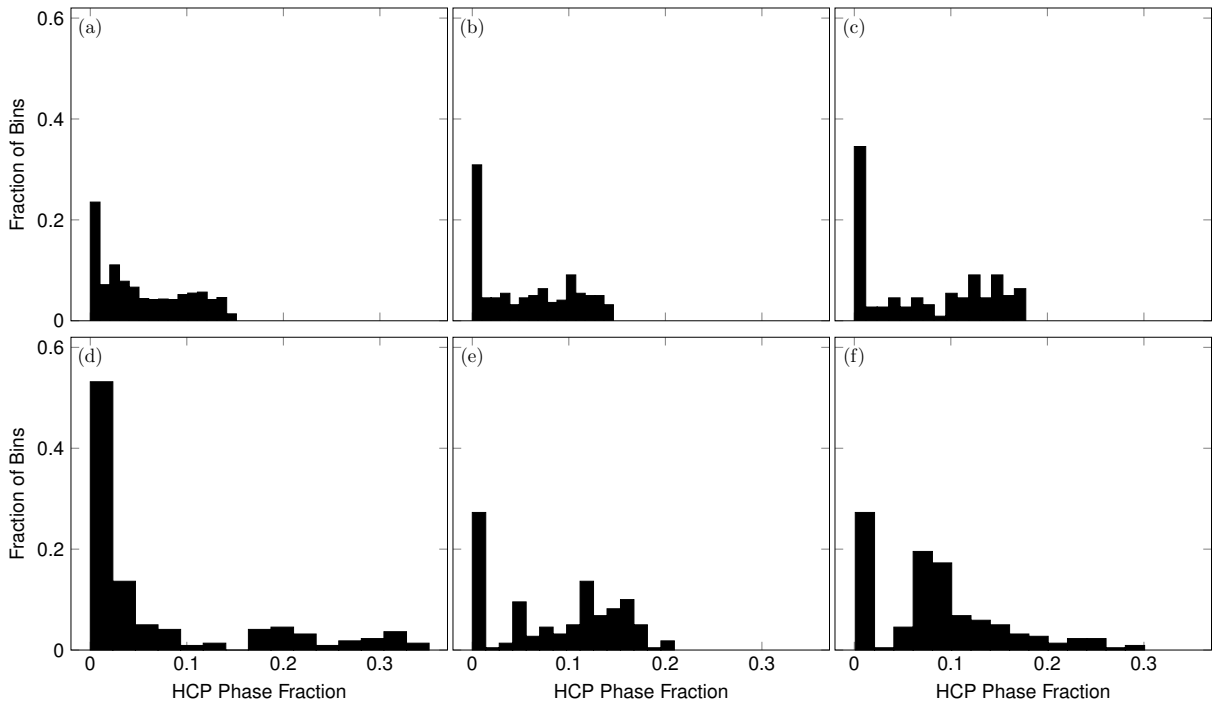

Fig. S5 Histograms of the HCP phase fractions measured in the binned regions for set numbers: a) 1, b) 2, c) 3, d) 4, e) 5, f) 11.

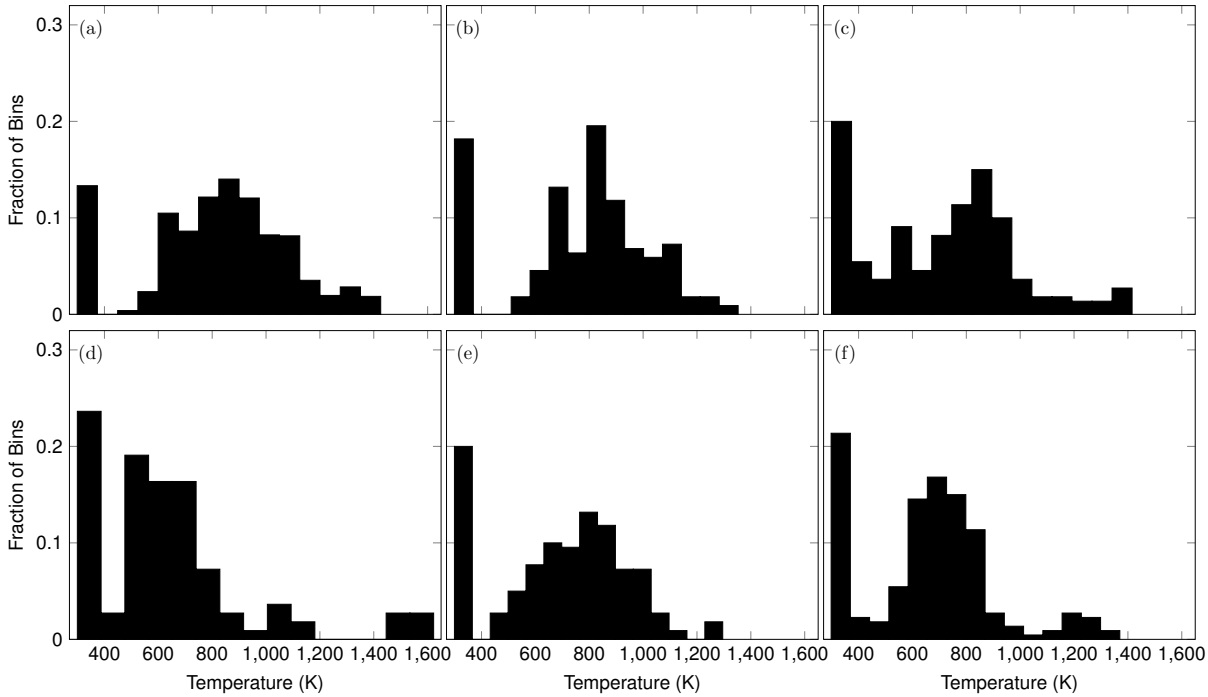

Fig. S6 Histograms of the temperatures measured in the binned regions for set numbers: a) 1, b) 2, c) 3, d) 4, e) 5, f) 11.

S3 Validation set regression accuracies for models trained with individual single-crystal shock-loading simulations

Tables S3, S4, S5, S6, and S7 provide the validation-set regression accuracy metrics for models trained with individual single-crystal shock-loading simulations. All metrics shown are MSE values computed using the portion of the corresponding dataset that was set aside for model validation and was not used for model training. MSE values were computed with the microstructure state descriptors being normalized to have values between 0 and 1. The “Average MSE" column shown in the tables provides the average MSE and its standard deviation computed for that microstructure state descriptor across the 11 models that were trained with different test-train splits of the corresponding dataset. The “Best Model MSE" column provides the MSE values for each descriptor from the model that had the lowest average MSE for the six microstructure state descriptors. The “Best Overall MSE" column shows the lowest MSE observed for that microstructure state descriptor across the 11 different models.

Examining the metrics provided in Tables S3, S4, S5, S6, and S7, we observe that models trained on individual single-crystal shock-loading simulations achieve similar prediction accuracy for the six microstructure state descriptors in their respective validation sets. Generally, the models are most accurate when predicting the pressure and the phase fractions of the disordered or FCC phases, and tend to have lower accuracy when predicting the total dislocation density and HCP phase fraction. The prediction accuracy for the temperature does vary depending on which shock-loading orientation was used to train the models, with the models trained with the $\langle 100 \rangle$ and $\langle 112 \rangle$ loading orientation simulations having temperature prediction accuracies similar to their respective phase fraction prediction accuracies, while the other three shock-loading orientations had temperature prediction accuracies which were noticeably worse. Examining the average and standard deviations for the MSE provided in the tables, it can be noted that the creation of the test-train split does have a significant impact on the performance of the models that are trained with those datasets. Comparisons between the MSE's from the best models to the best overall MSE's, it can be noted that the models with the lowest average MSE would not have the best MSE for every microstructure state descriptor, and at times the model with the lowest average MSE would have individual MSE values for specific microstructure state descriptors that were above average compared to the other models trained with that same data.

Table S3 Validation set MSE metrics for models trained with the complete $\langle 111 \rangle$ shock-loading simulation

Microstructure Descriptor	Average MSE	Best Model MSE	Best Overall MSE
Pressure	$7.16 \pm 5.25 \times 10^{-4}$	4.37×10^{-4}	3.80×10^{-4}
Total Dislocation Density	$1.79 \pm 0.633 \times 10^{-3}$	1.14×10^{-3}	1.09×10^{-3}
Disordered Fraction	$1.09 \pm 0.411 \times 10^{-3}$	5.10×10^{-3}	5.10×10^{-4}
FCC Fraction	$1.32 \pm 0.586 \times 10^{-3}$	5.77×10^{-4}	5.77×10^{-4}
HCP Fraction	$4.11 \pm 1.71 \times 10^{-3}$	2.53×10^{-3}	2.18×10^{-3}
Temperature	$1.54 \pm 0.433 \times 10^{-3}$	1.11×10^{-3}	8.41×10^{-4}

Table S4 Validation set MSE metrics for models trained with the truncated $\langle 111 \rangle$ shock-loading simulation

Microstructure Descriptor	Average MSE	Best Model MSE	Best Overall MSE
Pressure	$8.17 \pm 5.50 \times 10^{-4}$	2.60×10^{-4}	2.60×10^{-4}
Total Dislocation Density	$1.97 \pm 0.596 \times 10^{-3}$	1.03×10^{-3}	1.03×10^{-3}
Disordered Fraction	$5.32 \pm 2.99 \times 10^{-4}$	3.31×10^{-4}	2.60×10^{-4}
FCC Fraction	$6.62 \pm 5.36 \times 10^{-4}$	3.42×10^{-4}	2.33×10^{-4}
HCP Fraction	$2.24 \pm 1.18 \times 10^{-3}$	1.79×10^{-3}	1.38×10^{-3}
Temperature	$1.34 \pm 0.594 \times 10^{-3}$	1.08×10^{-3}	7.08×10^{-4}

Table S5 Validation set MSE metrics for models trained with the $\langle 110 \rangle$ shock-loading simulation

Microstructural Descriptor	Average MSE	Best Model MSE	Best Overall MSE
Pressure	$4.27 \pm 2.49 \times 10^{-4}$	2.28×10^{-4}	2.28×10^{-4}
Total Dislocation Density	$1.26 \pm 1.05 \times 10^{-3}$	5.36×10^{-4}	5.36×10^{-4}
Disordered Fraction	$8.14 \pm 8.35 \times 10^{-4}$	2.24×10^{-4}	2.24×10^{-4}
FCC Fraction	$5.20 \pm 3.82 \times 10^{-4}$	2.66×10^{-4}	2.02×10^{-4}
HCP Fraction	$2.14 \pm 1.33 \times 10^{-3}$	1.09×10^{-3}	1.00×10^{-3}
Temperature	$1.18 \pm 0.771 \times 10^{-3}$	1.26×10^{-3}	6.76×10^{-4}

Table S6 Validation set MSE metrics for models trained with the ⟨100⟩ shock-loading simulation

Microstructural Descriptor	Average MSE	Best Model MSE	Best Overall MSE
Pressure	$4.51 \pm 1.97 \times 10^{-4}$	2.94×10^{-4}	2.40×10^{-4}
Total Dislocation Density	$2.80 \pm 0.707 \times 10^{-3}$	1.78×10^{-3}	1.78×10^{-3}
Disordered Fraction	$6.23 \pm 2.54 \times 10^{-4}$	3.56×10^{-4}	3.56×10^{-4}
FCC Fraction	$6.20 \pm 2.07 \times 10^{-4}$	3.93×10^{-4}	3.93×10^{-4}
HCP Fraction	$8.08 \pm 2.76 \times 10^{-4}$	6.73×10^{-4}	4.17×10^{-4}
Temperature	$1.00 \pm 0.585 \times 10^{-3}$	6.21×10^{-4}	4.33×10^{-4}

Table S7 Validation set MSE metrics for models trained with the ⟨112⟩ shock-loading simulation

Microstructural Descriptor	Average MSE	Best Model MSE	Best Overall MSE
Pressure	$3.38 \pm 3.64 \times 10^{-4}$	1.33×10^{-4}	1.11×10^{-4}
Total Dislocation Density	$1.72 \pm 0.706 \times 10^{-3}$	9.53×10^{-4}	9.53×10^{-4}
Disordered Fraction	$6.12 \pm 6.08 \times 10^{-4}$	3.84×10^{-4}	2.87×10^{-4}
FCC Fraction	$5.62 \pm 8.06 \times 10^{-4}$	2.29×10^{-4}	2.16×10^{-4}
HCP Fraction	$1.21 \pm 0.794 \times 10^{-3}$	6.48×10^{-4}	5.37×10^{-4}
Temperature	$8.61 \pm 6.63 \times 10^{-4}$	4.70×10^{-4}	3.55×10^{-4}

S4 Best overall MSE for models trained with individual single-crystal shock-loading orientations and making predictions for different loading orientations

Tables S8, S9, S10, and S11 summarize the transferability of models trained with individual shock-loading orientation simulations by providing the best overall MSE observed when predicting the microstructure state descriptors for shock-loading orientations that were excluded from model training. As discussed in Section 3.3 of the main text, the ability of models trained with data from one shock-loading orientation to make predictions with data from a different shock-loading orientation depends on several factors. When the expected bounds for the microstructure state descriptors are significantly different between the two datasets, the model will tend to under or over-predict the descriptor depending on whether the training dataset had lower or higher bounds than the targeted dataset. This is most prominently observed for models predicting the descriptors from the ⟨100⟩ loading orientation dataset or for the predictions made by the models trained with the ⟨100⟩ loading orientation dataset, as the best overall MSEs observed for these tasks tend to be higher than the other MSE's reported in the tables.

Table S8 Best overall MSE for the models trained with the truncated ⟨111⟩ shock-loading simulation predicting descriptors for other loading orientation simulations

Microstructure Descriptor	⟨110⟩ Best Overall MSE	⟨100⟩ Best Overall MSE	⟨112⟩ Best Overall MSE
Pressure	4.34×10^{-3}	1.50×10^{-2}	4.42×10^{-3}
Total Dislocation Density	6.43×10^{-3}	1.01×10^{-2}	4.17×10^{-3}
Disordered Fraction	5.17×10^{-3}	8.57×10^{-3}	3.53×10^{-3}
FCC Fraction	3.52×10^{-3}	1.11×10^{-2}	4.25×10^{-3}
HCP Fraction	4.36×10^{-3}	1.27×10^{-2}	5.26×10^{-3}
Temperature	2.28×10^{-3}	8.50×10^{-3}	4.07×10^{-3}

Table S9 Best overall MSE for the models trained with ⟨110⟩ shock-loading simulation predicting descriptors for other loading orientation simulations

Microstructure Descriptor	⟨111⟩ Best Overall MSE	⟨100⟩ Best Overall MSE	⟨112⟩ Best Overall MSE
Pressure	9.81×10^{-3}	3.12×10^{-3}	2.87×10^{-3}
Total Dislocation Density	1.08×10^{-2}	3.45×10^{-2}	1.84×10^{-2}
Disordered Fraction	1.97×10^{-3}	2.11×10^{-2}	5.32×10^{-3}
FCC Fraction	3.53×10^{-3}	1.05×10^{-2}	6.25×10^{-3}
HCP Fraction	2.08×10^{-2}	1.26×10^{-1}	3.61×10^{-2}
Temperature	1.55×10^{-2}	1.13×10^{-2}	8.09×10^{-3}

Table S10 Best overall MSE for the models trained with ⟨100⟩ shock-loading simulation predicting descriptors for other loading orientation simulations

Microstructure Descriptor	⟨111⟩ Best Overall MSE	⟨110⟩ Best Overall MSE	⟨112⟩ Best Overall MSE
Pressure	1.13×10^{-2}	2.72×10^{-3}	2.92×10^{-3}
Total Dislocation Density	3.51×10^{-1}	5.55×10^{-1}	3.91×10^{-1}
Disordered Fraction	5.38×10^{-3}	1.60×10^{-2}	3.57×10^{-3}
FCC Fraction	4.51×10^{-3}	1.32×10^{-2}	6.55×10^{-3}
HCP Fraction	2.37×10^{-2}	2.32×10^{-2}	3.67×10^{-2}
Temperature	1.88×10^{-2}	2.23×10^{-2}	2.02×10^{-2}

Table S11 Best overall MSE for the models trained with ⟨112⟩ shock-loading simulation predicting descriptors for other loading orientation simulations

Microstructure Descriptor	⟨111⟩ Best Overall MSE	⟨110⟩ Best Overall MSE	⟨100⟩ Best Overall MSE
Pressure	1.28×10^{-2}	5.99×10^{-3}	5.06×10^{-3}
Total Dislocation Density	8.83×10^{-3}	3.10×10^{-2}	5.74×10^{-2}
Disordered Fraction	4.11×10^{-3}	6.44×10^{-3}	1.31×10^{-2}
FCC Fraction	2.33×10^{-3}	3.26×10^{-3}	6.85×10^{-3}
HCP Fraction	1.22×10^{-2}	1.87×10^{-2}	3.81×10^{-2}
Temperature	1.83×10^{-2}	3.12×10^{-2}	6.23×10^{-2}

S5 Best overall MSE for models trained with three single-crystal shock-loading orientations and making predictions for the excluded single crystal simulation

Table S12 provides the best overall MSE's observed for models that were trained with three single crystal shock-loading orientation simulations when predicting the microstructure state descriptors with data from the single crystal simulation that was excluded from model training. In general, metrics reported in Table S12 are lower than those reported in the tables from Supplementary Information S4, indicating that using multiple shock-loading orientation simulations to train the models improves their transferability to unseen shock-loading orientations.

Table S12 Best overall MSE observed for models trained with three single crystal shock-loading orientations and predicting on the orientation that was excluded from training

Microstructure Descriptor	⟨111⟩ Best Overall MSE	⟨110⟩ Best Overall MSE	⟨100⟩ Best Overall MSE	⟨112⟩ Best Overall MSE
Pressure	4.37×10^{-3}	1.52×10^{-3}	3.06×10^{-3}	7.88×10^{-4}
Total Dislocation Density	3.27×10^{-3}	1.62×10^{-2}	2.71×10^{-2}	3.23×10^{-3}
Disordered Fraction	1.72×10^{-3}	3.36×10^{-3}	1.33×10^{-2}	2.79×10^{-3}
FCC Fraction	1.37×10^{-3}	1.74×10^{-3}	4.74×10^{-3}	2.16×10^{-3}
HCP Fraction	4.40×10^{-3}	5.46×10^{-3}	3.37×10^{-2}	3.55×10^{-3}
Temperature	1.09×10^{-2}	6.38×10^{-3}	1.97×10^{-2}	8.23×10^{-3}

S6 Best overall MSE for models trained with data from single crystal shock-loading simulations and predicting descriptors for the polycrystalline simulation

Tables S13 and S14 summarize the transferability of models trained with data from single crystal shock-loading simulations for predicting microstructure state descriptors from polycrystalline shock simulations. The dataset numbers shown in the tables indicate the dataset used to train the models, with the numbers referring to the datasets defined in Table S1. The metrics shown for dataset 11 in Table S13 are from models trained with the polycrystalline data making predictions with the validation portion of the polycrystalline dataset, which serve as best-possible performances to compare the models trained with single crystal data against.

As discussed in Section 3.4 of the main text, some microstructure state descriptors such as temperature and total dislocation density aren't predicted well by any of the of the models trained with single crystal data. Others, such as pressure and phase fractions, can be predicted with accuracies comparable to those observed in the other model transfer cases shown in Tables S8, S9, S10, S11, and S12.

Table S13 Best overall MSE observed for models trained with individual single crystal shock-loading simulations and predicting on the polycrystalline shock simulation

Microstructure Descriptor	Dataset 11 (validation) Best Overall MSE	Dataset 2 Best Overall MSE	Dataset 3 Best Overall MSE	Dataset 4 Best Overall MSE	Dataset 5 Best Overall MSE
Pressure	3.94×10^{-4}	3.52×10^{-3}	2.89×10^{-3}	1.44×10^{-3}	4.44×10^{-3}
Total Dislocation Density	3.00×10^{-3}	6.47×10^{-2}	5.97×10^{-2}	4.90×10^{-2}	5.66×10^{-2}
Disordered Fraction	1.23×10^{-3}	2.70×10^{-3}	5.02×10^{-3}	1.04×10^{-2}	2.78×10^{-3}
FCC Fraction	6.78×10^{-4}	3.08×10^{-3}	3.38×10^{-3}	4.24×10^{-3}	2.30×10^{-3}
HCP Fraction	1.36×10^{-3}	4.75×10^{-2}	2.61×10^{-2}	2.38×10^{-2}	2.05×10^{-2}
Temperature	1.71×10^{-3}	3.65×10^{-2}	1.78×10^{-2}	8.99×10^{-3}	3.49×10^{-2}

Table S14 Best overall MSE observed for models trained with multiple single crystal shock-loading simulations and predicting on the polycrystalline shock simulation

Microstructure Descriptor	Dataset 6 Best Overall MSE	Dataset 7 Best Overall MSE	Dataset 8 Best Overall MSE	Dataset 9 Best Overall MSE	Dataset 10 Best Overall MSE
Pressure	1.31×10^{-3}	1.83×10^{-3}	1.90×10^{-3}	2.24×10^{-3}	1.62×10^{-3}
Total Dislocation Density	2.90×10^{-2}	3.21×10^{-2}	3.62×10^{-2}	4.44×10^{-2}	2.17×10^{-2}
Disordered Fraction	2.48×10^{-3}	2.10×10^{-3}	2.75×10^{-3}	2.82×10^{-3}	2.75×10^{-3}
FCC Fraction	1.36×10^{-3}	1.08×10^{-3}	1.56×10^{-3}	2.06×10^{-3}	1.00×10^{-3}
HCP Fraction	4.12×10^{-3}	4.40×10^{-3}	4.24×10^{-3}	1.40×10^{-2}	9.28×10^{-3}
Temperature	1.10×10^{-2}	1.03×10^{-2}	1.12×10^{-2}	1.31×10^{-2}	1.04×10^{-2}

S7 Comparisons with a Simpler Machine-Learning Approach

As a point of comparison to the machine-learning model architectures described in Supplementary Information 1, we present here the model training results for a simpler machine-learning workflow. In this simpler approach, the normalized XRD profiles are passed directly into a decision tree regression model to predict the microstructural state descriptors. For this series of tests, scikit-learn's decision tree model `GradientBoostingRegressor`[3,4] was used for the regression of the microstructural descriptors. As this implementation only allows for the regression of a single regression target, individual single-output regression models were trained for each of the six microstructural descriptors.

For this comparison, two training datasets were used: Dataset 2 (the truncated $\langle 111 \rangle$ dataset) and Dataset 7 (the combination of the truncated $\langle 111 \rangle$, $\langle 110 \rangle$, and $\langle 100 \rangle$ datasets). For each of the training datasets, ten different test-train splits of 70% training and 30% test were used to train ten different iterations of the direct regression workflow. For the models trained with Dataset 2, the trained models were used to predict the microstructural state descriptors from Datasets 3, 4, 5, and 11. The models trained with Dataset 7 were used to predict the microstructural state descriptors from Datasets 5 and 11.

S7.1 Direct Regression from XRD Profiles

For the direct regression workflow, normalized XRD profiles were passed directly to the gradient boosting decision tree regression model for the prediction of the microstructural state descriptors. Table S15 provides validation set regression accuracy metrics for the models trained with Dataset 2. Comparing the metrics provided in Table S15 with those provided in Table S4, we can observe that the simpler direct regression workflow does have worse regression accuracy metrics compared to the sequential autoencoder-MLP workflow described in Supplementary Information 1, with substantial differences in most MSE metrics, and slightly worse performance in the prediction of the disordered phase fraction by the most accurate direct regression model.

Table S15 Validation set MSE metrics for models trained with the truncated $\langle 111 \rangle$ shock-loading simulation

Microstructure Descriptor	Average MSE	Best Model MSE	Best Overall MSE
Pressure	$1.26 \pm 0.984 \times 10^{-3}$	6.88×10^{-4}	5.62×10^{-4}
Total Dislocation Density	$2.91 \pm 1.07 \times 10^{-3}$	1.14×10^{-3}	1.14×10^{-3}
Disordered Fraction	$7.92 \pm 2.22 \times 10^{-4}$	5.76×10^{-4}	5.37×10^{-4}
FCC Fraction	$7.23 \pm 5.18 \times 10^{-4}$	4.52×10^{-4}	4.00×10^{-4}
HCP Fraction	$6.72 \pm 2.37 \times 10^{-3}$	4.16×10^{-3}	3.82×10^{-3}
Temperature	$2.42 \pm 1.38 \times 10^{-3}$	1.74×10^{-3}	1.16×10^{-3}

Table S16 provides the lowest observed MSE across the ten different test-train splits for the direct regression models that were trained with Dataset 2 when predicting the microstructural state descriptors for Datasets 3, 4, 5, and 11. These metrics can be directly

compared to the accuracies provided in Tables S8 and S13. Comparing the ability of the direct regression models to make predictions on different loading orientation datasets compared to the performance of the more complex machine-learning workflow described in Supplementary Information 1, we can observe that for most of the microstructural descriptors, the more complex machine-learning workflow significantly outperforms the direct regression approach. However, for Datasets 3, 4, and 5, the direct regression approach does perform similarly or outperform the autoencoder-MLP workflow in predicting the disordered phase fraction and the FCC phase fraction. Also, for Dataset 11, the direct regression approach does have a lower MSE for the prediction of the total dislocation density, although both the direct regression approach and the autoencoder-MLP approach perform poorly for this descriptor.

Table S16 Best overall MSE observed for models trained with Dataset 2 and predicting on Datasets 3, 4, 5, and 11

Microstructure Descriptor	Dataset 3 Best Overall MSE	Dataset 4 Best Overall MSE	Dataset 5 Best Overall MSE	Dataset 11 Best Overall MSE
Pressure	1.17×10^{-2}	4.10×10^{-2}	1.21×10^{-2}	1.47×10^{-2}
Total Dislocation Density	2.31×10^{-2}	2.54×10^{-2}	1.03×10^{-2}	4.32×10^{-2}
Disordered Fraction	3.63×10^{-3}	5.53×10^{-3}	2.80×10^{-3}	3.71×10^{-3}
FCC Fraction	4.22×10^{-3}	6.18×10^{-3}	2.67×10^{-3}	4.31×10^{-3}
HCP Fraction	5.91×10^{-2}	2.38×10^{-1}	4.32×10^{-2}	8.13×10^{-2}
Temperature	2.86×10^{-2}	3.51×10^{-2}	1.90×10^{-2}	6.15×10^{-2}

Table S17 provides validation set regression accuracy metrics for direct regression models trained with Dataset 7, as well as prediction accuracy metrics when these models predict microstructural state descriptors for Datasets 5 and 11. Comparing the metrics provided in Table S17 with those provided in Tables S12 and S14, we can observe that the direct regression workflow has the same performance trade-offs when trained with data from multiple shock-loading orientations when predicting on different single crystal or polycrystalline shock-loading simulations. Predictions of the disordered phase fraction and FCC phase fraction are better for the direct regression workflow compared to the autoencoder-MLP workflow for Dataset 5, while predictions for the other microstructural state descriptors are worse. Predictions of the total dislocation density of Dataset 11 are slightly better but still poor with the direct regression workflow compared to the autoencoder-MLP workflow, while the regression accuracies of the other microstructrual state descriptors are better for the autoencoder-MLP workflow.

Table S17 MSE metrics for models trained with Dataset 7

Microstructure Descriptor	Dataset 7 (validation) Average MSE	Dataset 7 (validation) Best Overall MSE	Dataset 5 Best Overall MSE	Dataset 11 Best Overall MSE
Pressure	$9.85 \pm 7.85 \times 10^{-4}$	4.90×10^{-4}	2.16×10^{-3}	6.18×10^{-3}
Total Dislocation Density	$2.62 \pm 0.821 \times 10^{-3}$	1.30×10^{-3}	5.08×10^{-3}	2.62×10^{-2}
Disordered Fraction	$6.39 \pm 1.07 \times 10^{-4}$	4.69×10^{-4}	1.18×10^{-3}	2.97×10^{-3}
FCC Fraction	$6.54 \pm 0.912 \times 10^{-4}$	4.86×10^{-4}	1.19×10^{-3}	2.37×10^{-3}
HCP Fraction	$1.48 \pm 0.317 \times 10^{-3}$	9.21×10^{-4}	7.96×10^{-3}	4.59×10^{-3}
Temperature	$2.35 \pm 0.541 \times 10^{-3}$	1.68×10^{-3}	9.38×10^{-3}	2.28×10^{-2}

In summary, the direct regression workflow is capable of learning the relationships that exist between XRD profiles collected during shock-loading simulations and sets of microstructural state descriptors. Direct regression models achieve relatively good (although worse than the autoencoder-MLP workflow) regression accuracies for validation sets. With respect to the transferability of direct regression models to data from shock-loading orientations that were excluded from the training data, there is a consistent trend where the direct regression approach performs better than the autoencoder-MLP workflow when predicting the disordered phase fractions and FCC phase fractions of unseen single crystal shock-loading simulations, while performing worse (in some cases, such as predicting the pressure, substantially worse) on all of the other tested microstructural state descriptors. When predicting the microstructural state descriptors from the polycrystalline shock simulation, the direct regression approach does consistently do better when predicting the total dislocation density compared to the autoencoder-MLP workflow (however, both perform poorly for this particular descriptor), while performing worse for all of the other descriptors. In general, the direct regression approach performs worse than the autoencoder-MLP workflow defined in Supplementary Information 1, and the cases where the direct regression approach performs better than the autoencoder-MLP workflow typically come at the cost of a substantial decrease in performance in other areas that compromise its utility.

S8 PCA Analysis of XRD Data from Molecular Dynamics Simulations of the Shock Loading of Cu

Principal Component Analysis (PCA)[5] can be used to estimate how many latent dimensions will be needed to accurately encode a dataset by examining the amount of variance that each additional latent dimension captures. Figure S7 provides the amount of variance captured by each latent dimension up to a latent dimension size of 20, as well as the cumulative variance that is captured as the number of latent dimensions is increased. For this analysis, the PCA implementation from sckikit-learn was used with all of the xrd profiles from Dataset 6.

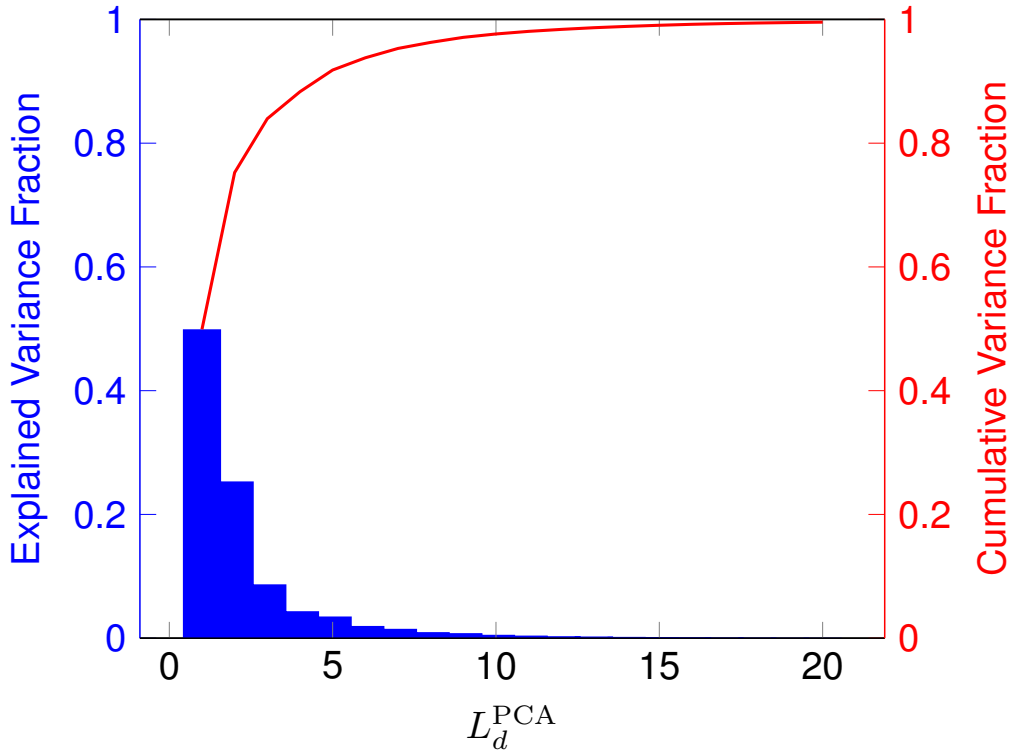

Fig. S7 Explained variance and cumulative variance as a function of the intrinsic dimensionality, L_d. Note that $L_d = 20$ captures over 99.5% of the variance in our dataset.

Per the analysis of the variance captured as a function of latent dimension size, a latent dimension of size 20 captures over 99.5% of the variance present in the XRD data collected from molecular dynamics simulations of the shock loading of single crystal Cu along different loading orientations. With this analysis as well as a sensitivity study that was performed by examining the reconstruction performance of autoencoders with the XRD data as a function of encoded dimension size, a latent dimension of size 20 was selected for our machine-learning workflows.

S9 Shock Compression versus Spall Failure

Figure S8 compares the distributions of the stress, temperature, and dislocation density along the loading axis of the simulation cell for the four different single crystal configurations during (a) shock compression and (b) spall failure. These comparisons visualize how varying the loading orientation changes the evolution of the microstructural descriptors at equivalent stages of the shock simulation.

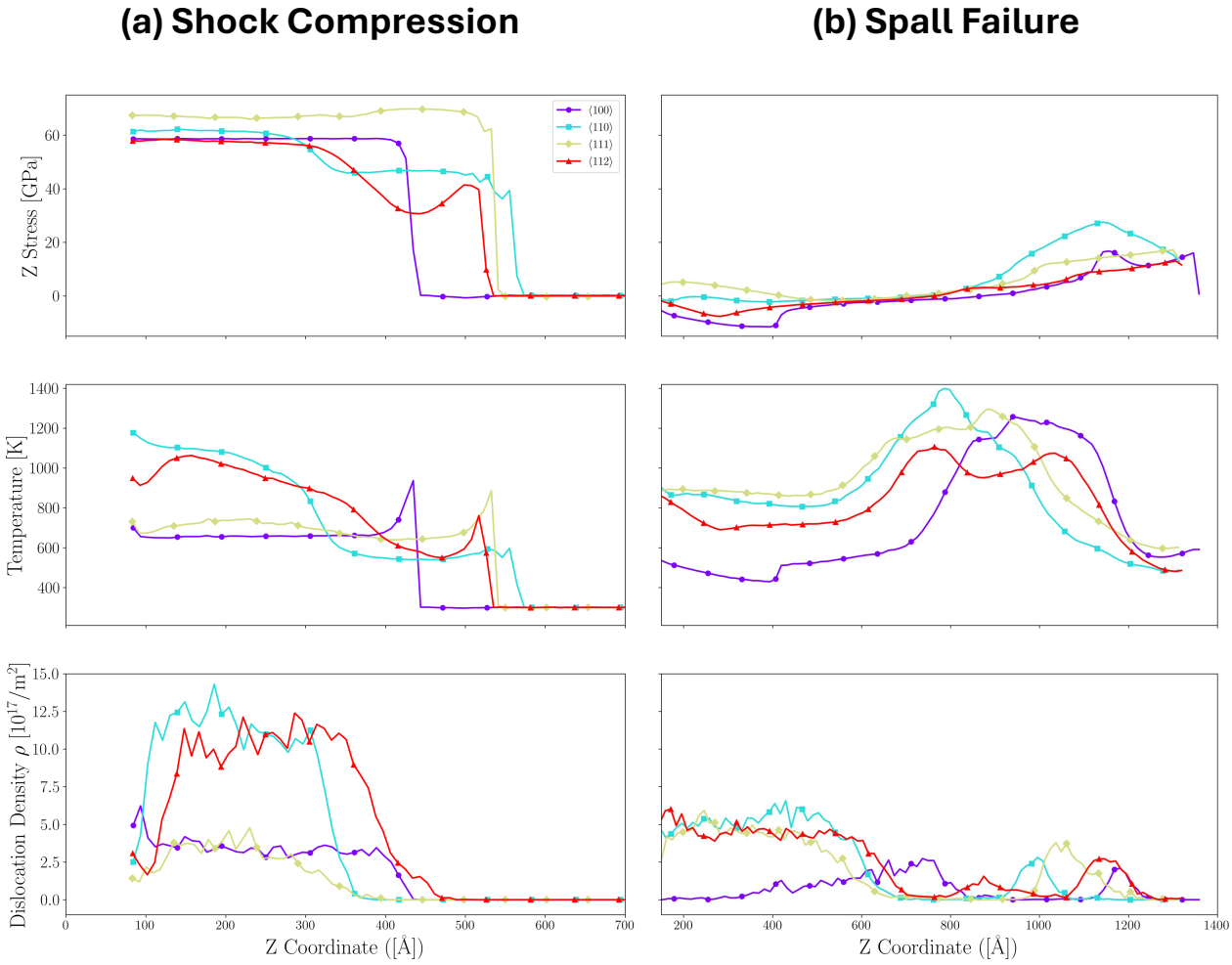

Fig. S8 Z stress, temperature, and dislocation density versus position along the shock-loading direction for different shock-loading orientation simulations during (a) shock compression and (b) spall failure.

Supplementary References

1 D. Vizoso, G. Subhash, K. Rajan and R. Dingreville, *Chemistry of Materials*, 2013, **35**, 1186–1200.

2 D. Vizoso and R. Dingreville, *Journal of Applied Physics*, 2025, **137**, XX.

3 J. H. Friedman, *The Annals of Statistics*, 2001, **29**, 1189–1232.

4 J. H. Friedman, *Computational Statistics & Data Analysis*, 2002, **38**, 367–378.

5 A. Mead, *Journal of the Royal Statistical Society*, 1992, **41**, 27–39.