Supplementary Information

MELT-3D: Zero-shot cross-material prediction of three-dimensional full-field crystal-plasticity evolution with a latent transformer

This supplement provides implementation details omitted from the main text (Sections S1–S4) and the extended orientation analysis summarized in the Discussion (Section S5).

S1 Data, simulations, and field conventions

S1.1 Training exposure

Table S1: Training exposure of each component to the evaluation data.

Component	Training data	Cu	HEA-Si	HEA-Ni
Transformer	64^3 CP trajectories, Cu + HEA-Si	held-out sims	held-out sims	zero-shot
Shared codec	64^3 CP trajectories, Cu + HEA-Si	held-out sims	held-out sims	zero-shot
Rotation-only codec	128^3 synthetic, undeformed	unseen	unseen	unseen

Table S1 summarizes which components of the pipeline were exposed to each material. HEA-Ni is zero-shot for every learned component. The rotation-only codec, which is a diagnostic model not coupled to the latent transformer (Section S5), was trained on undeformed synthetic polycrystals and has therefore seen none of the three evaluated materials and no deformed state.

S1.2 Constitutive parameters

The cubic elastic constants and Voce hardening parameters of the three material systems are listed in Table S2 in the form used by the condition vector, together with the position of each HEA-Ni value relative to the range spanned by the two training materials.

S1.3 Dataset allocation

Partitioning was performed at the simulation level before individual deformation states and spatial patches were extracted, so no patch from a test simulation appears in any training batch. Of the 150 simulations available for each of Cu and HEA-Si, 135 were assigned to the training set, 5 to the validation set, and 10 to the held-out test set. The 40 HEA-Ni simulations form an independent zero-shot pool and were never used for training, validation, early stopping, or hyperparameter selection. All held-out and zero-shot simulations were available for evaluation; the number of simulations included in each analysis is stated in the corresponding figure or table caption.

Table S2: Material descriptors and the position of HEA-Ni relative to the training range. Interpolation indicates that the HEA-Ni value lies within the range spanned by Cu and HEA-Si for that descriptor considered alone; extrapolation indicates that it lies outside. Because the descriptors are evaluated componentwise, interpolation in every component would not by itself imply that the complete condition vector lies within the convex hull of the training set.

Descriptor	Cu	HEA-Si	HEA-Ni	HEA-Ni relative to training range
C_{11} (MPa)	168400	249000	172000	Interpolation
C_{12}/C_{11}	0.7209	0.6265	0.6279	Interpolation
C_{44}/C_{11}	0.4477	0.5703	0.5349	Interpolation
τ_0 (MPa)	45	120	102	Interpolation
$\tau_0 + \tau_1$ (MPa)	75	320	422	Extrapolation
θ_0 (MPa)	800	650	500	Extrapolation
θ_1 (MPa)	130	240	40	Extrapolation

S1.4 Field channel ordering and conventions

Each recorded state comprises 24 channels: six components giving the first two columns of the crystallographic rotation matrix R, followed by six elastic-strain, six plastic-strain, and six stress components. The strain and stress components are stored in Voigt order, $(11, 22, 33, 23, 13, 12)$, with the shear strains stored as tensor components ε_{ij} rather than as engineering components $\gamma_{ij} = 2\varepsilon_{ij}$. Strains and stresses therefore share the same component ordering and the same factor convention, and no conversion is applied at any stage of the framework. The third column of R is recovered as the cross product of the first two and is therefore not stored. The rotation components are dimensionless and bounded in $[-1, 1]$; strains are dimensionless and stresses are in MPa.

The spectral solver produces no zero-strain output, so the first recorded state corresponds to an applied strain of 2×10^{-4} rather than to the undeformed configuration. This state is supplied to the model as the initial condition, and every subsequent recorded state is a prediction target compared with the crystal-plasticity field at the same applied strain. The first predicted state is therefore at 0.0004, one increment beyond the supplied initial condition.

The Cu simulations comprise 49 recorded states and the HEA-Si and HEA-Ni simulations 300, at a constant increment of $\Delta\varepsilon = 2 \times 10^{-4}$, giving 48 and 299 prediction targets and final applied strains of 0.0098 and 0.0600, respectively. The Cu trajectory was limited to approximately 1% strain because this range captured the elastic regime, the elastic–plastic transition, and the subsequent plastic-deformation regime; restricting it reduced the data volume and computational cost while retaining all three deformation regimes.

S2 Condition vector and zero-shot normalization

S2.1 Definitions and scaling of the condition-vector components

Deformation progress. The applied strain of the requested state, divided by a fixed reference strain of $\varepsilon_{\mathrm{ref}} = 0.10$. The variable therefore spans 0.004–0.098 over the Cu trajectory and 0.004–0.600 over the longer HEA-Si and HEA-Ni trajectories, and takes the value 0.002 at the supplied initial condition. The normalizer is a fixed reference strain rather than the maximum applied strain of each trajectory, so a given value denotes the same absolute deformation for every material; normalizing per trajectory would have made the conditioning variable depend on the extent of the

Table S3: Components of the 26-dimensional transformer condition vector. The same vector is supplied to the statistics predictor. No categorical material label is included.

Components	Dimension	Description
Deformation progress	1	Normalized target deformation step
Loading descriptors	18	Nine masked loading components and a nine-component mask
Elastic descriptors	3	Normalized cubic elastic constants
Plasticity descriptors	4	Voce hardening parameters
Total	26	

target-material simulation and would therefore have been inadmissible for zero-shot operation. The reference strain exceeds the largest applied strain in the training set, so no training state lies at the upper end of the variable's range.

Loading descriptors. The imposed loading is described by nine components of the applied velocity-gradient or stress boundary condition together with a nine-component binary mask indicating which components are prescribed. Under the strain-controlled uniaxial tension considered here the mask is identical across all simulations, so these 18 entries are constant within the present dataset; they are retained so that the same condition vector can describe multiaxial and mixed boundary conditions without architectural change.

Elastic descriptors. The three cubic elastic constants enter as C_{11}/C_{ref}, C_{12}/C_{11}, and C_{44}/C_{11}, with $C_{\mathrm{ref}} = 200{,}000$ MPa. The two ratios are dimensionless measures of the relative stiffness and of the elastic anisotropy, so the material is described by one magnitude and two shape parameters rather than by three correlated stiffnesses.

Plasticity descriptors. The four Voce parameters enter as τ_0 and $\tau_0 + \tau_1$, each divided by a reference stress of 200 MPa, and θ_0 and θ_1, each divided by a reference stress of 1000 MPa. Two normalizers are used because slip resistances and hardening moduli differ systematically in magnitude, so a single scale would leave the hardening descriptors with a much larger dynamic range than the resistances. The saturation stress is supplied as the sum $\tau_0 + \tau_1$ rather than as τ_1 alone so that each descriptor corresponds directly to a feature of the flow curve: the initial slip resistance, the saturation resistance, the initial hardening rate, and the asymptotic hardening rate.

S2.2 Voce hardening law and the mechanics baseline

The Voce flow stress used to construct the mechanics baseline is

$$\sigma_f(\varepsilon_p) = M_T \left[\tau_0 + (\tau_1 + \theta_1\Gamma)\left(1 - \exp\left(-\frac{\theta_0\Gamma}{\tau_1}\right)\right)\right], \qquad \Gamma = M_T\,\varepsilon_p, \tag{S1}$$

where Γ is the accumulated slip, $M_T = 2.528$, determined by least squares against the volume-averaged axial stresses of the Cu and HEA-Si training simulations. The fitted value is treated thereafter as a fixed constant of the mechanics baseline: it is not re-fit for, and does not use any data from, the zero-shot material. Because the baseline supplies only an initial estimate of the channel statistics, which the learned correction then adjusts, a single scalar is adequate; systematic departures arising from texture or from the full-constraint assumption are absorbed by

that correction. The single-crystal parameters are those of Table S2. The elastic constants are homogenized by Voigt–Reuss–Hill averaging to give K, G, E, and ν, and the uniaxial response follows from the consistency condition

$$\varepsilon_p = \max\Big(\varepsilon_t - \frac{\sigma}{E},\, 0\Big), \qquad \sigma = \sigma_f(\varepsilon_p), \tag{S2}$$

solved by fixed-point iteration to a relative tolerance of 10^{-8} at each deformation state.

The resulting uniaxial solution supplies the 18 channel means $\hat{\mu}_{\mathrm{mech}}$. For loading along the z direction, representative components are

$$\hat{\mu}_{\mathrm{mech}}[\sigma_{zz}] = \sigma, \qquad \hat{\mu}_{\mathrm{mech}}[\varepsilon^e_{xx}] = -\nu\varepsilon_e, \qquad \hat{\mu}_{\mathrm{mech}}[\varepsilon^p_{xx}] = -\tfrac{1}{2}\varepsilon_p, \tag{S3}$$

with $\varepsilon_e = \sigma/E$; the transverse plastic components follow from plastic incompressibility and the off-diagonal means are zero under uniaxial loading. The channel-specific scales s_{mech} are constructed as fixed fractions of the corresponding characteristic magnitude — the axial stress for the stress channels, ε_e for the elastic-strain channels, and ε_p for the plastic-strain channels — so that each scale grows with the field it normalizes. A single fixed Taylor factor is adequate because the mechanics model supplies only a baseline estimate; systematic departures arising from texture or from the full-constraint assumption are absorbed by the learned correction.

S2.3 Floor values and statistics-predictor training

Lower bounds are imposed on s_{mech} and on the entries of σ_{floor} to prevent division by vanishingly small scales during the earliest elastic states, where several plastic-strain channels are identically zero and the corresponding empirical standard deviation is numerically indistinguishable from zero. The floors are 10^{-6} for the strain channels and 0.1 MPa for the stress channels. Both are several orders of magnitude below the corresponding characteristic scales at any state beyond the first few increments, so they are active only in the early elastic regime and do not affect the reported results.

For Cu and HEA-Si, the training targets are

$$y_{\mathrm{resid}} = (\mu_{\mathrm{emp}} - \hat{\mu}_{\mathrm{mech}}) \oslash r_{\mathrm{ref}}, \qquad y_{\log\sigma} = \log\left[\max(\sigma_{\mathrm{emp}}, \sigma_{\mathrm{floor}}) \oslash s_{\mathrm{mech}}\right], \tag{S4}$$

where $\oslash$ denotes element-wise division and the maximum and logarithm are evaluated element-wise. The reference-scale vector r_{ref} contains 10^{-2} for the strain channels and 100 MPa for the stress channels, so that the residual corrections to the means are of order unity for both field types. The residual head is zero-initialized, so the untrained network reproduces the mechanics baseline exactly, and the log-scale head has its output bias initialized to the training-set mean of the log-standard-deviation targets. Training used Adam with an initial learning rate of 10^{-3}, learning-rate reduction on a validation-loss plateau, and early stopping; every fifth deformation state from each training material was withheld for validation. The empirical statistics used to form these targets were calculated using only simulations assigned to the training set.

S3 Neural-network architectures and training

S3.1 Variational autoencoder

All VAE convolutions use unit stride and same padding; consequently, the spatial resolution remains 8×8×8 throughout the encoder and decoder. Batch normalization and the Swish activation function are applied after the convolutional layers unless otherwise noted.

Table S4: Architecture of the VAE encoder. Here, C3 denotes a $3 \times 3 \times 3$ convolution and C1 denotes a $1 \times 1 \times 1$ convolution.

Stage	Input	Operation
Orientation branch	Six orientation channels	C3(512), followed by four C3(256) layers
Elastic branch	Six elastic-strain channels	Five C3(256) layers
Plastic branch	Concatenated elastic- and plastic-strain channels	Five C3(256) layers
Stress branch	Concatenated elastic-strain and stress channels	Five C3(256) layers
Feature fusion	Four 256-channel branch outputs	Concatenation to 1024 channels
Residual block 1	1024-channel fused field	Four C3(256) layers with a C1(256) skip projection
Residual blocks 2–3	256-channel field	Four C3(256) layers per block with identity skip connections
Attention gate	256-channel field	C1(256) with sigmoid activation, followed by element-wise multiplication
Latent-mean tower	Gated 256-channel field	C3(256), C3(128), C3(64), and linear C3(1)
Latent-log-variance tower	Gated 256-channel field	C3(256), C3(128), C3(64), and linear C3(1)

The four reconstruction terms of Eq. (6) of the main text are equally weighted. $\mathcal{L}_{\mathrm{rot}}$ is the mean geodesic angle between the reference and reconstructed rotation matrices, evaluated without minimization over the cubic symmetry operators during training; cubic-symmetry-reduced disorientation is used subsequently for crystallographic evaluation. The mechanical reconstruction terms are Huber losses with transition parameter $\delta = 0.4$. The Kullback–Leibler term weakly regularizes the approximate posterior toward a unit normal distribution; its small weight (10^{-6}) reflects that the latent field is used here as a compressed representation for downstream prediction rather than as a generative prior.

S3.2 Conditioned latent transformer

The initial $8 \times 8 \times 8$ latent field is flattened into 512 scalar spatial tokens. A dense projection maps each scalar to a d_{model}-dimensional embedding. Three trainable axial positional-embedding tables of size $8 \times d_{\mathrm{model}}$ represent the three spatial directions; the position assigned to a spatial token is the sum of the corresponding entries from the three tables.

The deformation-progress variable is embedded using a three-layer multilayer perceptron with an initial hidden dimension of 256. The loading, elasticity, and plasticity groups are each embedded using two-layer multilayer perceptrons with a hidden dimension of 128. The elastic, loading, and plasticity embeddings form three additional tokens that are prepended to the spatial sequence.

Each transformer block uses pre-normalized multi-head self-attention followed by a two-layer feed-forward network with GELU activation. Adaptive scale, shift, and residual-gating parameters are generated separately for the attention and feed-forward sublayers, and all adaptive parameters are zero-initialized. Dropout is applied to the attention output, the intermediate feed-forward representation, and the feed-forward output. After the transformer blocks the condition tokens are removed and the spatial tokens are mapped to a scalar latent increment through a pointwise

Table S5: Architecture of the VAE decoder. Supplying the predicted elastic-strain field to the plastic-strain and stress heads mirrors the corresponding concatenations in the encoder and retains coupling among the mechanical fields.

Stage	Input	Operation
Shared trunk	$8 \times 8 \times 8 \times 1$ latent field	C3(512), three 512-channel residual blocks, and two C3(256) layers
Orientation head	Shared 256-channel field	C3 layers with channel widths $(256, 128, 32, 32, 32, 16, 12)$, followed by linear C3(6) and Gram–Schmidt orthonormalization
Elastic-strain head	Shared 256-channel field	C3 layers with widths $(64, 64, 32, 32, 32)$, a sigmoid gate, C3 layers with widths $(16, 12)$, linear C3(6), and a second sigmoid gate
Plastic-strain head	Shared field concatenated with the predicted elastic strain	C3 layers with widths $(128, 64, 32, 32, 32, 16, 12)$, followed by linear C3(6)
Stress head	Shared field concatenated with the predicted elastic strain	C1 layers with widths $(128, 64, 32, 32, 32, 16, 12)$, followed by linear C1(6)

multilayer-perceptron head. The final output layer is zero-initialized, so the untrained network reproduces the identity mapping $\hat{z}_t = z_0$ exactly.

S3.3 Orientation-only diagnostic codec

The orientation-only codec supports the diagnostic reported in Section S5. It was not coupled to the latent transformer and does not alter any of the delivered field predictions reported in the main text. It shares the latent geometry of the shared codec — an $8 \times 8 \times 8 \times 1$ field, unit stride and same padding throughout — but is an independent design rather than the shared encoder branch and decoder head with the other three fields removed.

Three input operations have no counterpart in the shared codec. The six input channels are orthonormalized by Gram–Schmidt, so the encoder operates on a valid pair of rotation columns; the axis-angle logarithm of the resulting rotation matrix is concatenated, giving nine channels and an explicit representation of rotation magnitude and axis; and normalized voxel coordinates are concatenated after the first pointwise convolutions.

The encoder then applies C1(64) and C1(32) pointwise layers, coordinate augmentation, a C3(64) stem, two residual blocks — the first retaining 64 channels with an identity skip, the second widening to 96 with a dilated first convolution and a C1(96) skip projection — and two C3(128) layers, followed by linear C3(1) towers for the latent mean and log-variance. The decoder mirrors this: C3(128), the same two residual blocks, a C3(64) and C3(32) tail, two C1(64) generator layers, and a linear C1(6) output followed by Gram–Schmidt orthonormalization. Batch normalization and the Swish activation follow each convolution. The dilated convolutions enlarge the receptive field within the 8^3 patch without downsampling, and the terminal Gram–Schmidt guarantees a valid pair of orthonormal columns, from which the third follows by cross product as for the shared codec.

The reconstruction loss is the mean geodesic angle between reference and reconstructed rotation

Table S6: Architecture and hyperparameters of the conditioned latent transformer.

Parameter	Value
Initial latent-field shape	$8 \times 8 \times 8$
Number of spatial tokens	512
Condition-vector dimension	26
Number of condition tokens	3
Total sequence length	515
Token-embedding dimension	384
Axial positional embeddings	Three 8×384 tables
Number of transformer blocks	8
Number of attention heads	12
Dimension per attention head	32
Feed-forward hidden dimension	1536
Condition-embedding dimension	384
Dropout rate	0.10
Prediction-head dimensions	512 and 256
Conditioning mechanism	AdaLN-Zero and condition tokens
Output parameterization	$\hat{z}_t = z_0 + \Delta z_t$
Prediction mode	Direct, non-autoregressive
Optimizer	AdamW
Loss	Mean Huber loss
Learning-rate schedule	Linear warm-up followed by cosine decay
Gradient stabilization	Global gradient-norm clipping
Number of model parameters	Approximately 22.0 million

matrices, evaluated without minimization over the cubic symmetry operators. This is identical to the $\mathcal{L}_{\text{rot}}$ term of the shared VAE loss, so the two codecs are compared under the same objective and symmetry convention.

The codec contains approximately 3.2 million trainable parameters in total, against approximately 8.9 million in the orientation branch of the shared encoder alone (Table S4). The dedicated codec is therefore the smaller model, and its advantage cannot be attributed to greater capacity devoted to orientation. Training used 8^3 patches extracted from undeformed 128^3 synthetic polycrystals generated with DREAM.3D.

S4 Evaluation implementation details

S4.1 Metric definitions and error decomposition

All predictions are evaluated after decoding and transformation to physical units.

Macroscopic response. The macroscopic axial stress is the volume average of the predicted voxelwise stress over the reconstructed 64^3 volume,

$$\sigma_{33}(t) = \frac{1}{N^3} \sum_{\mathbf{r}} \sigma_{33}(\mathbf{r}, t). \tag{S5}$$

The corresponding relative error is calculated with respect to the crystal-plasticity result. Relative errors at very early deformation states are interpreted cautiously because the reference stress is small and a modest absolute discrepancy produces a large relative error; this accounts for the maxima reported in the main text, all of which occur in the early elastic regime.

Normalized mean absolute error. For a six-component field group, the voxelwise mean absolute error is calculated separately for each component and normalized by the spatial standard deviation of the corresponding reference component; the normalized component errors are then averaged over the six components. Normalizing by the spatial standard deviation rather than the mean is necessary for two reasons. Shear components have volume averages close to zero under uniaxial loading, so normalization by the mean would be ill-conditioned for them. And absolute stress levels differ substantially among Cu, HEA-Si, and HEA-Ni, so an error measure normalized by field variation is required for the three materials to be compared on a common scale.

Separation of compression and prediction errors. Two reconstruction pathways separate the error introduced by latent compression from that introduced by conditioned latent prediction. The decoded reference latent is $X_t^{\mathrm{codec}} = D_\psi(z_t)$, whereas the complete prediction is $X_t^{\mathrm{pred}} = D_\psi(\hat{z}_t)$. The compression error compares X_t^{codec} with the original field X_t; the additional decoded prediction discrepancy compares X_t^{pred} with X_t^{codec}; the end-to-end error compares X_t^{pred} directly with X_t. Because the decoder D_ψ acts on both latent fields, the second comparison isolates the effect of latent prediction after passage through a common decoder. It does not remove the decoder's nonlinear sensitivity, so the standalone prediction discrepancy should not be read as an additive component of the end-to-end error. For crystallographic orientation this non-additivity is exact rather than approximate, because disorientations are geodesic distances on $SO(3)$ subject to the triangle inequality.

Crystallographic orientation. For reference and predicted rotation matrices R and $\hat{R}$, the voxelwise disorientation is

$$\theta = \min_{S \in G_{\mathrm{cubic}}} \cos^{-1}\left(\frac{\mathrm{tr}\left(SR\hat{R}^T\right) - 1}{2}\right), \tag{S6}$$

where G_{cubic} is the proper cubic symmetry group. Without symmetry reduction, crystallographically equivalent orientations would be assigned angular differences of up to the 62.8° maximum disorientation for cubic symmetry. Kernel average misorientation is calculated from the symmetry-reduced disorientation between each voxel and its 26 nearest neighbors.

Reported uncertainties. Unless otherwise stated, reported curves and uncertainties represent the mean and standard deviation over independent simulations, with the number of simulations given in each figure or table caption. Where a quantity is averaged over deformation states rather than simulations, the caption states the number of states and the uncertainty is the standard deviation over those states.

S4.2 Spatial autocorrelation and boundary ratios

The correlation length ℓ of an error field is the lag at which its normalized spatial autocorrelation first falls below $1/e$. The autocorrelation is computed by FFT on each two-dimensional slice and

averaged over all 64 planes. The boundary-to-interior ratio br is the mean absolute error at voxels lying on an 8^3 patch face divided by the mean absolute error at patch-interior voxels, so $br = 1$ indicates no dependence on position within the patch. Subscripts denote the error source: codec is $\mathrm{decode}(z_{\mathrm{true}}) - \mathrm{raw}$ and model is $\mathrm{decode}(z_{\mathrm{pred}}) - \mathrm{decode}(z_{\mathrm{true}})$.

Table S7: Spatial error structure, by material and field, averaged over five sampled deformation states spanning the elastic–plastic transition and subsequent hardening. ℓ_{codec} and ℓ_{model} are correlation lengths (voxels) of the codec and model error fields, respectively; br_{codec} and br_{model} are boundary-to-interior error ratios; excess is $br_{\mathrm{model}} - br_{\mathrm{codec}}$.

Material	Field	ℓ_{codec}	ℓ_{model}	br_{codec}	br_{model}	Excess
Cu	r_{33}	1.00	1.00	1.092	0.996	−0.097
Cu	ε^e_{33}	2.00	3.00	0.990	1.119	+0.128
Cu	ε^p_{33}	2.00	4.00	1.083	1.012	−0.071
Cu	σ_{33}	1.00	3.00	1.058	1.213	+0.154
HEA-Si	r_{33}	1.00	1.00	1.090	0.999	−0.091
HEA-Si	ε^e_{33}	2.00	3.00	0.993	1.125	+0.131
HEA-Si	ε^p_{33}	2.00	4.00	1.097	1.020	−0.077
HEA-Si	σ_{33}	1.00	3.40	1.077	1.201	+0.124
HEA-Ni	r_{33}	1.00	1.00	1.094	1.020	−0.074
HEA-Ni	ε^e_{33}	2.00	3.20	0.992	1.079	+0.086
HEA-Ni	ε^p_{33}	1.80	3.60	1.096	1.039	−0.057
HEA-Ni	σ_{33}	1.00	3.20	1.079	1.141	+0.062

The excess column isolates the transformer's boundary dependence from whatever positional dependence the shared per-patch decode already imposes. That baseline is itself field-dependent: the codec error is mildly elevated at patch faces for r_{33}, ε^p_{33}, and σ_{33} ($br_{\mathrm{codec}} = 1.06$–$1.10$) but marginally suppressed for ε^e_{33} (0.990–0.993), so the subtraction matters in both directions.

Correlation lengths separate the two error sources for every mechanical field: codec error is short ranged ($\ell = 1.0$–2.0) while transformer error is coherent over 3.0–4.0 voxels, approximately grain scale at this resolution. Rotation is the exception, with $\ell = 1.0$ for both, indicating that the transformer adds no spatially coherent structure to the orientation error.

S4.3 Latent-residual direction controls

The latent drift ratio reported in the main text shows that the predicted state does not move as far from the supplied initial state as the encoded reference state. We performed an additional diagnostic to determine whether the corresponding prediction residual lies in a latent direction to which the decoder is unusually insensitive.

For each evaluated state, the observed latent residual was defined as $r_{\mathrm{obs}} = z_{\mathrm{true}} - z_{\mathrm{pred}}$, so that applying it to the predicted latent state recovers the encoded reference state. Two control residuals were constructed. The random control r_{rand} was drawn isotropically and rescaled to have the same Euclidean norm as the observed residual. The shuffled control r_{shuf} was obtained by permuting the components of the observed residual, which preserves the component values and Euclidean norm while destroying their spatial arrangement in the latent field.

For a residual r applied at the predicted latent state, its decoded-field cost was calculated as

$$C(r; z_{\mathrm{pred}}) = d_{\mathrm{field}}[D_\psi(z_{\mathrm{pred}} + r),\, D_\psi(z_{\mathrm{pred}})]\,, \tag{S7}$$

where d_{field} is the normalized decoded-field discrepancy used for this diagnostic. The same field components, channel normalization, and spatial averaging were used for the observed and control residuals, so the cost ratios $R_{\text{rand}} = C(r_{\text{obs}})/C(r_{\text{rand}})$ and $R_{\text{shuf}} = C(r_{\text{obs}})/C(r_{\text{shuf}})$ compare the effects of residual direction at fixed latent-space magnitude. A ratio of unity indicates that the observed and control residuals produce equal decoded-field perturbations.

Table S8: Decoded-field cost of the observed latent residual relative to norm-matched random and component-shuffled control residuals. Values are averaged over the second half of each trajectory (ε beyond 50% of the sampled deformation range; $n = 5$ states per material). Uncertainties denote one standard deviation over these states.

Quantity	Cu	HEA-Si	HEA-Ni
Observed/random residual	1.0481 ± 0.0033	1.0520 ± 0.0020	1.0565 ± 0.0035
Observed/shuffled residual	1.0046 ± 0.0012	1.0020 ± 0.0015	1.0017 ± 0.0010

The observed-to-random cost ratio was 1.048–1.057, so the observed residual produced a decoded-field perturbation approximately 5% larger than a norm-matched random residual. The observed-to-shuffled ratio was 1.002–1.005, indicating nearly identical decoded-field costs for the observed residual and its component-shuffled counterpart. Both comparisons were similar for Cu, HEA-Si, and HEA-Ni. These results provide no evidence that the observed prediction residual is concentrated in an unusually decoder-insensitive direction. The diagnostic rules out one possible explanation for the difference between the latent drift ratio and the decoded-field errors. It does not imply that latent-space distances and decoded-field errors should be numerically proportional, because the two quantities use different definitions and normalizations.

S5 Extended orientation analysis

This section provides the complete state-by-state orientation error decomposition summarized in the main text (Section S5.1), and supports the dedicated-codec diagnostic discussed under Future directions (Sections S5.2–S5.5).

S5.1 Complete decomposition of the orientation error

Table S9 reports the codec, transformer, and end-to-end disorientation statistics at every evaluated deformation state, extending the representative states given in the main text. The codec establishes the larger and more stable contribution for all three materials, and at matched strain the HEA-Si and HEA-Ni codec medians differ by no more than approximately 0.08°.

Relation between the three measures. The codec and transformer disorientations cannot be added to obtain the end-to-end disorientation. At each voxel, the three orientations define geodesic distances on $SO(3)$ that satisfy $|\theta_{\text{codec}} - \theta_{\text{xf}}| \leq \theta_{\text{e2e}} \leq \theta_{\text{codec}} + \theta_{\text{xf}}$, and the reported medians are calculated independently over the voxels of each disorientation field, so the median end-to-end error is not expected to equal the sum of the codec and transformer medians. A practical measure of how much latent-state prediction increases the typical delivered error is the difference between the end-to-end and codec medians: at the final evaluated state this is 0.29° for Cu at $\varepsilon = 0.0098$, 0.71° for HEA-Si at $\varepsilon = 0.06$, and 0.72° for HEA-Ni at $\varepsilon = 0.06$.

S5.2 The dedicated orientation codec

The decomposition identifies compression as the binding constraint on orientation fidelity, but it does not establish whether that constraint is intrinsic to the six-component rotation representation or a consequence of sharing latent capacity across all 24 channels. To test this, we evaluated a separate codec trained only on the six rotation components (Section S3.3). This orientation-only codec is a diagnostic model: it is not coupled to the latent transformer and does not alter any result reported in the main text. Its purpose is to bound what a better representation could achieve.

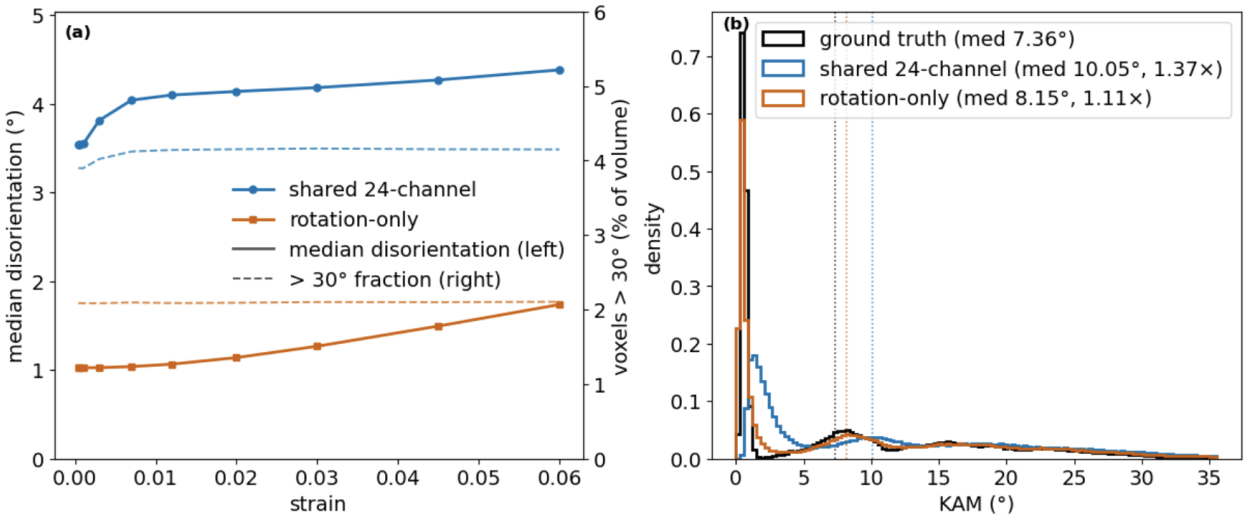


Figure S1: Comparison of the shared 24-channel codec and the orientation-only codec for HEA-Ni, which was absent from the training sets of both codecs. (a) Median symmetry-reduced disorientation as a function of applied strain for the two codecs (solid lines, left axis) and the fraction of voxels with disorientation greater than 30° (dashed lines, right axis). (b) Distribution of kernel average misorientation (KAM) at $\varepsilon = 0.030$ for the reference field and the two codec reconstructions. Dotted vertical lines indicate the median of each distribution; legend entries report the median and, for each reconstruction, its ratio to the reference value. Each state contains 262,144 voxels.

Across the evaluated strain range, the median disorientation of the shared codec increases from 3.55° to 4.38°, whereas that of the orientation-only codec ranges from 1.03° to 1.74° (Fig. S1a, Table S10) — a reduction of 2.52× to 3.89× depending on deformation state. The dedicated codec also reproduces the local orientation variation far more closely. At $\varepsilon = 0.030$, the median KAM is 7.36° for the reference field, 10.06° for the shared-codec reconstruction, and 8.15° for the orientation-only reconstruction (Fig. S1b), so the reconstructed-to-reference KAM ratio falls from 1.37 to 1.11. The dedicated reconstruction recovers both features of the reference distribution: the sharp grain-interior peak below 1°, which the shared codec smears across several degrees, and the broader boundary population near 8°.

The gain is concentrated in the bulk of the distribution rather than its tail. The median improves by up to 3.9×, but the mean and 90th percentile improve by only 2.0–2.2×, and the fraction of voxels exceeding 30° is merely halved. Summary statistics dominated by the tail — the mean in particular — therefore understate the improvement by nearly a factor of two.

We report HEA-Ni here because it is zero-shot for every component of the pipeline, so the comparison is made on a composition neither codec has seen. The two codecs are nonetheless not equally exposed even here: the shared codec was trained on 64^3 crystal-plasticity trajectories of other materials and has therefore seen deformed microstructures, whereas the dedicated codec

was trained only on undeformed synthetic polycrystals and has seen neither this material nor any deformed state. Cu and HEA-Si behave similarly (Fig. S2); for those materials the asymmetry is larger still, since the shared codec was trained on their own trajectories while the dedicated codec remains unexposed.

S5.3 Controlling for differences in the training distribution

The shared and dedicated codecs differ in both the fields they represent and the data used for training. The dedicated codec was trained on 8^3 patches extracted from 128^3 undeformed synthetic polycrystals. In contrast, the shared codec was trained on 8^3 patches extracted from 64^3 crystal-plasticity trajectories similar to those used to evaluate both codecs. Training at 128^3 was computationally feasible because generating synthetic polycrystals is considerably less expensive than performing crystal-plasticity simulations at the same resolution.

Consequently, spatial resolution and deformation-state coverage differ simultaneously between the two training sets. At 128^3, each grain is represented by more voxels, and a fixed 8^3 patch generally covers a smaller portion of the microstructure. However, the grain interiors in these undeformed synthetic structures have uniform orientations by construction. The dedicated codec was therefore trained on higher-resolution orientation fields but was not exposed to the intragranular orientation gradients that develop during deformation. The shared codec was trained at a lower spatial resolution but had direct exposure to these deformation-induced features.

Because both codecs process fixed 8^3 patches, the global resolution of the training and evaluation volumes does not require a change in architecture. Nevertheless, the differences in the training distributions introduce a potential confounding factor: the improved performance of the dedicated codec may result partly from the smoother, higher-resolution fields used for training, rather than solely from allocating its latent capacity to orientation.

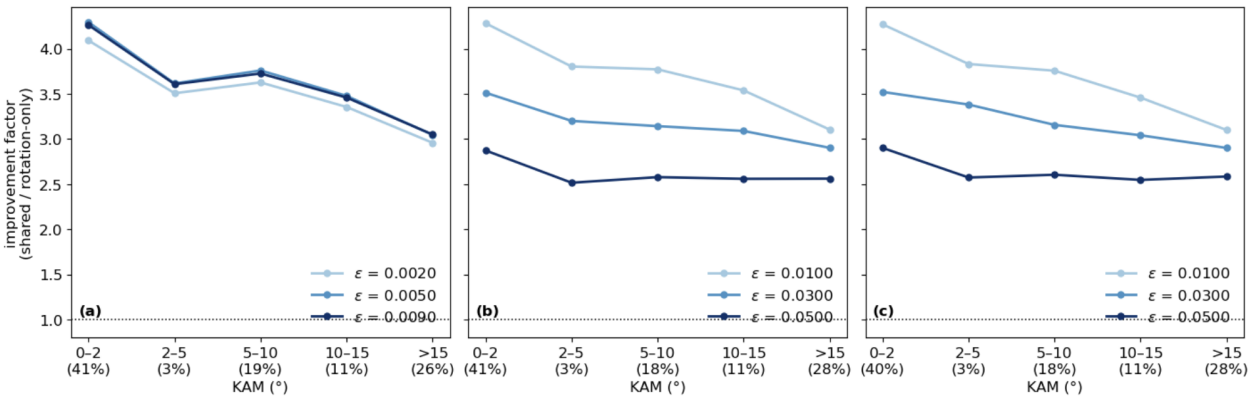


Figure S2: The advantage of the dedicated codec persists at matched local difficulty, for every material. Improvement factor — the ratio of shared to dedicated median disorientation — resolved by ground-truth kernel average misorientation, at three strains per material: (a) Cu, (b) HEA-Si, (c) HEA-Ni. Binning by ground-truth KAM places both codecs at matched local orientation spread, the quantity that distinguishes their training sets. The volume fraction of each bin is given beneath the axis; the three microstructures are statistically equivalent in gradient distribution. HEA-Si is represented in the shared codec's training set and HEA-Ni in no training set, yet their curves coincide to within 0.03× throughout. Copper, simulated only to $\varepsilon = 0.0098$, shows no strain dependence over its shorter trajectory. The dotted line marks parity.

To reduce the influence of differences in local field complexity, the voxels are grouped according

to the local orientation gradient in the reference field, and the two codecs are compared within each bin (Fig. S2). The improvement factor ranges from 2.5× to 4.3× across all gradient bins, deformation states, and materials. For the two high-entropy alloys, the improvement factor decreases from the lowest-gradient bin (0–2°) to the highest-gradient bin (> 15°) by 27–28% at $\varepsilon = 0.010$, 17–18% at $\varepsilon = 0.030$, and 11% at $\varepsilon = 0.050$. Thus, the dedicated codec retains a substantial advantage over the full range of local orientation gradients, although the advantage is somewhat smaller in regions with large gradients. Because the dedicated codec is evaluated at a lower spatial resolution than that used for its training, these results may also underestimate its performance at its native training resolution.

The binned comparison further indicates that the improvement is not specific to a particular material. For HEA-Si and HEA-Ni, the improvement factors differ by less than 0.03× in every gradient bin and at every matched strain. This agreement is observed even though HEA-Si was included in the shared codec's training data, whereas HEA-Ni was absent from both training sets. The fractions of voxels in the five gradient bins also differ by less than one percentage point among the three materials, indicating that their distributions of local orientation gradients are closely matched. Copper, which yields at a lower strain and was simulated only to $\varepsilon = 0.0098$, shows no measurable decrease in the improvement factor over its shorter deformation trajectory. At a matched strain of approximately $\varepsilon = 0.01$, all three materials have improvement factors of 4.26–4.28× in the lowest-gradient bin. These results indicate that the subsequent decrease in the improvement factor is associated with the magnitude of accumulated deformation rather than with material identity or the relative position within each simulated trajectory.

The improvement factor does not vary monotonically with the local orientation gradient. The 2–5° bin contains only approximately 3% of the voxels in each material and lies between the grain-interior and grain-boundary populations shown in Fig. S1(b). The improvement factor decreases in this sparsely populated bin and then increases in the more populated 5–10° bin. This nonmonotonic variation does not affect the main result: the dedicated codec reduces the median error by at least a factor of 2.5 at every evaluated level of local orientation gradient.

S5.4 Residual high-angle errors and decreasing improvement with deformation

The remaining reconstruction errors are characterized by two observations: a persistent population of high-angle errors and a decrease in the relative improvement of the dedicated codec with increasing deformation.

The residual high-angle errors are not controlled primarily by codec fidelity. For the dedicated codec, the fraction of voxels with errors greater than 30° changes by less than 1% over the deformation trajectory. Similarly, the error survival curve above approximately 10° changes by less than 2% between $\varepsilon = 0.0004$ and $\varepsilon = 0.060$, even though the median disorientation increases by 69% over the same interval. Except at the first deformation state, the shared codec shows the same behavior. Thus, deformation increases the reconstruction error for most voxels but has little effect on the population of high-angle errors. This behavior is observed for both codecs and along the independent decoding paths.

Several results indicate that these high-angle errors correspond to orientations that are not recovered, rather than to small perturbations of the correct orientations. First, the reconstructed fields are locally rougher than the reference fields, with KAM values that are 1.31–1.44× higher. This behavior is inconsistent with simple interpolation across grain boundaries, which would instead smooth the orientation field. Second, the error distribution is nearly uniform between 30° and 57° and terminates at 61.7°, close to the maximum disorientation of 62.8° under cubic symmetry. Its

quantiles are also within 2–3° of those obtained when each voxel is paired with a random orientation drawn from the same texture.

The high-error voxels also have little correspondence with the orientations in their local neighborhoods. Their median disorientation from the closest of the 26 neighboring reference orientations is 27.0°, compared with 27.4° for the random null model. A small fraction of the errors can be attributed to assignment to an adjacent grain: 9.8% of the failed voxels are within 10° of a reference orientation in the immediate neighborhood, compared with 4.3% for the random null model. This excess arises entirely from nearest-neighbor matches. After accounting for the random expectation, approximately one in twenty failures can therefore be associated with an incorrect grain assignment. For the remaining failures, the reconstructed orientation is not related to either the correct orientation or those of neighboring grains.

These failures occur preferentially in regions with large orientation gradients. For the shared codec, their mean local reference KAM is 2.4–2.6× the volume-averaged value. Their prevalence is therefore more consistent with limited spatial resolution than with insufficient latent capacity. At a resolution of 64^3, some grains occupy only a few voxels, while the median KAM of the reference fields is approximately 7°. Under these conditions, small grains and sharp local orientation variations may not be adequately resolved. Increasing latent capacity alone is therefore unlikely to eliminate this error population.

Allocating a dedicated latent representation to orientation reduces the number of high-error voxels but does not change their statistical character. The failures of the two codecs also do not form a simple subset. At $\varepsilon = 0.060$, the shared and dedicated codecs fail at 10,885 and 5,518 voxels, respectively, with 2,768 failed voxels common to both. The dedicated codec corrects 8,117 failures of the shared codec but introduces 2,750 failures at different locations. Some of the remaining errors are associated with patch boundaries rather than with the orientation representation itself. At the same local orientation gradient, voxels on the faces of the 8^3 patches fail at 1.3× the rate of voxels in the patch interiors.

The relative improvement decreases with increasing deformation. The improvement factor decreases from 3.89× at $\varepsilon = 0.0070$ to 2.52× at $\varepsilon = 0.060$. This decrease is caused primarily by the increasing reconstruction error of the dedicated codec. Over this strain range, its median error increases by 67% (1.04° → 1.74°), whereas the median error of the shared codec increases by only 8% (4.04° → 4.38°). The most direct explanation is the difference in the training data. The dedicated codec was trained only on undeformed synthetic polycrystals, whose grain interiors have uniform orientations by construction. It was therefore not trained to represent the intragranular orientation gradients that develop during deformation. In contrast, these features were present in the training data used for the shared codec.

Consistent with this explanation, the reconstruction error increases as deformation-induced intragranular orientation gradients develop. Plastic deformation introduces lattice curvature associated with geometrically necessary dislocations, and the median KAM of the reference fields increases from 6.81° to 7.90° over the evaluated strain range. The strongest degradation occurs in the lowest-gradient bin, which represents nominal grain interiors and accounts for approximately 40% of the volume. Between $\varepsilon = 0.010$ and 0.050, the median error of the dedicated codec in this bin increases by 56% (0.69° → 1.09°), compared with an increase of only 6% for the shared codec (2.96° → 3.15°).

Model capacity may also contribute to this behavior. A codec operating near its reconstruction-error floor may show a clear increase in error when additional orientation structure develops, whereas this increase may be less apparent in a codec whose error is already dominated by limited

capacity. The effects of training coverage and model capacity cannot be separated using the present results. In either case, the dedicated codec is evaluated on deformation-induced structures absent from its training data. Because the dedicated codec was trained only on undeformed microstructures, the reported improvement factors—particularly at higher strains—likely underestimate the improvement that could be achieved by training the dedicated codec on both undeformed and deformed microstructures.

S5.5 Ways to improving orientation representation

The orientation representation could be improved by addressing three limitations of the current approach. First, the dedicated orientation only codec reported here was trained only on initial, undeformed microstructures. Including structures sampled throughout the deformation trajectory would expose the codec to the intragranular orientation gradients that develop during loading. Second, the remaining high-error voxels are primarily associated with limited spatial resolution rather than insufficient compression capacity. Training and evaluation at a finer resolution would provide more voxels per grain and should improve the representation of small grains and localized orientation gradients. Generating such high-resolution microstructures is also substantially less expensive than performing crystal-plasticity simulations at the same resolution. Third, voxels located on patch faces fail at 1.3× the rate of interior voxels with comparable orientation gradients. Overlapping patches during encoding, or adding a halo region similar to that proposed for the transformer, could reduce these boundary-related errors.

Expanding the training data to include deformed microstructures does not require a complete set of 128^3 crystal-plasticity simulations. Three lower-cost approaches could be evaluated. First, the dedicated codec could be fine-tuned using the existing 64^3 deformation trajectories. This would trade spatial resolution for direct coverage of deformation-induced intragranular structure and would test whether such coverage is more important than the resolution advantage of the 128^3 initial microstructures. Second, synthetic intragranular orientation gradients could be added to the 128^3 microstructures. These gradients could be constructed as smooth orientation fields consistent with geometrically necessary dislocation content, with their magnitudes selected to reproduce the kernel average misorientation (KAM) distribution observed in the crystal-plasticity trajectories. Third, the synthetic training set could be supplemented with a small number of 128^3 crystal-plasticity simulations. This approach assumes that the codec requires a representative sample of deformed intragranular structures rather than a complete ensemble of deformation trajectories.

Comparing these approaches would help determine how much of the remaining 1.0–1.7° error arises from limited training-data coverage and how much arises from model capacity. In addition, because the orientation-only codec was evaluated independently of the transformer, the improvement retained after integration into the complete prediction pipeline remains to be determined.

Table S9: Disorientation statistics for the codec, transformer, and end-to-end error measures. Codec denotes the disorientation between the reference orientation field and decode(z_{true}). Transformer denotes the decoded discrepancy between decode(z_{true}) and decode(z_{pred}); using the same decoder for both latent states isolates the effect of latent-state prediction after decoding but does not remove decoder sensitivity. End-to-end denotes the disorientation between the reference field and decode(z_{pred}). Statistics are calculated over all 262,144 voxels in each volume, and all angles are disorientations minimized over cubic crystal symmetry.

			Codec			Transformer			End-to-end	
Material	State	ε	Mean	Med.	p90	Mean	Med.	p90	Med.	p90
Cu	1	0.0004	6.262	3.504	12.593	0.407	0.251	0.498	3.528	12.602
Cu	2	0.0006	6.266	3.507	12.613	0.694	0.427	0.882	3.556	12.615
Cu	4	0.0010	6.358	3.582	12.796	1.831	1.050	2.738	3.683	12.785
Cu	6	0.0014	6.488	3.698	12.966	2.757	1.642	4.290	3.913	13.170
Cu	14	0.0030	6.738	3.911	13.396	3.674	2.242	5.978	4.204	13.722
Cu	20	0.0042	6.791	3.959	13.513	3.823	2.340	6.257	4.252	13.820
Cu	27	0.0056	6.820	3.987	13.572	3.903	2.396	6.414	4.278	13.879
Cu	34	0.0070	6.836	4.003	13.608	3.948	2.429	6.502	4.293	13.899
Cu	41	0.0084	6.846	4.013	13.639	3.975	2.447	6.553	4.305	13.914
Cu	48	0.0098	6.855	4.023	13.661	3.994	2.464	6.583	4.314	13.926
HEA-Si	1	0.0004	6.362	3.541	12.868	0.062	0.034	0.092	3.543	12.876
HEA-Si	2	0.0006	6.362	3.543	12.874	0.388	0.241	0.474	3.561	12.886
HEA-Si	4	0.0010	6.374	3.550	12.887	0.774	0.453	1.050	3.594	12.952
HEA-Si	14	0.0030	6.732	3.891	13.553	3.518	2.151	5.686	4.179	13.967
HEA-Si	34	0.0070	6.909	4.058	13.932	4.031	2.498	6.675	4.371	14.260
HEA-Si	59	0.0120	6.947	4.102	14.000	4.114	2.548	6.850	4.419	14.298
HEA-Si	99	0.0200	6.969	4.136	13.994	4.165	2.586	6.934	4.481	14.317
HEA-Si	159	0.0320	7.009	4.192	14.047	4.254	2.667	7.036	4.614	14.366
HEA-Si	224	0.0450	7.074	4.269	14.078	4.386	2.804	7.195	4.807	14.459
HEA-Si	299	0.0600	7.166	4.382	14.161	4.584	3.006	7.468	5.087	14.627
HEA-Ni[†]	1[‡]	0.0004	6.381	3.546	13.062	3.919	2.809	5.492	4.398	14.227
HEA-Ni[†]	2	0.0006	6.381	3.546	13.053	0.320	0.199	0.400	3.562	13.075
HEA-Ni[†]	4	0.0010	6.381	3.547	13.058	0.446	0.277	0.575	3.570	13.084
HEA-Ni[†]	14	0.0030	6.669	3.813	13.676	3.256	1.963	5.238	4.066	13.953
HEA-Ni[†]	34	0.0070	6.911	4.041	14.116	4.132	2.537	6.877	4.386	14.560
HEA-Ni[†]	59	0.0120	6.962	4.099	14.185	4.289	2.650	7.179	4.470	14.707
HEA-Ni[†]	99	0.0200	6.996	4.139	14.243	4.350	2.710	7.284	4.540	14.738
HEA-Ni[†]	159	0.0320	7.038	4.193	14.285	4.410	2.779	7.360	4.657	14.770
HEA-Ni[†]	224	0.0450	7.099	4.268	14.321	4.517	2.904	7.495	4.836	14.827
HEA-Ni[†]	299	0.0600	7.190	4.382	14.382	4.688	3.099	7.720	5.104	14.964

[†]Zero-shot: HEA-Ni was not seen during training.

[‡]The first predicted HEA-Ni state is anomalous: its median decoded transformer discrepancy is 2.81°, an order of magnitude larger than the 0.20° obtained at the next state, and the elevation is confined to the transformer term while the codec term is unaffected. The behavior is reproduced across all ten HEA-Ni microstructures and is absent for Cu and HEA-Si. Because it occurs only at the smallest deformation-progress value p_t and only for the material whose hardening descriptors lie outside the training range, it is consistent either with the conditioning pathway behaving anomalously as $p_t \to 0$ or with the learned correction extrapolating poorly at the first state; we have not determined which. The state is excluded from the ranges quoted in the main text to describe the subsequent evolution.

Table S10: Disorientation statistics for the shared 24-channel and dedicated rotation-only codecs on HEA-Ni. The improvement factor is the ratio of shared to dedicated median disorientation. All angles in degrees, over 262,144 voxels per state.

	Shared 24-channel			Rotation-only			
ε	Med.	Mean	p90	Med.	Mean	p90	Improv.
0.0004	3.55	6.38	13.06	1.03	3.13	6.43	3.46×
0.0010	3.55	6.38	13.06	1.03	3.13	6.44	3.46×
0.0030	3.81	6.67	13.68	1.03	3.13	6.43	3.71×
0.0070	4.04	6.91	14.12	1.04	3.14	6.44	3.89×
0.0120	4.10	6.96	14.18	1.07	3.16	6.45	3.85×
0.0200	4.14	7.00	14.24	1.14	3.22	6.47	3.63×
0.0300	4.18	7.03	14.27	1.27	3.31	6.51	3.30×
0.0450	4.27	7.10	14.32	1.49	3.48	6.59	2.86×
0.0600	4.38	7.19	14.38	1.74	3.67	6.73	2.52×