

Supplementary Material for Trustworthy Machine Learning of Diffractograms to Prevent Silent Failure of Apparently Accurate Models

Redad Mehdi^{1,3}, Weiqi Yue^{2,3}, Ozan Dernek^{1,3}, Matthew A. Willard^{1,3}, Roger H.
French^{1,2,3}, and Frank Ernst^{1,3,*}

¹Case Western Reserve University, Department of Materials Science and Engineering,
Cleveland, OH 44106, USA

²Case Western Reserve University, Department of Computer and Data Sciences, Cleveland,
OH 44106, USA

³Case Western Reserve University, Materials Data Science for Stockpile Stewardship,
Cleveland, OH 44106, USA

*fxe5@case.edu

The baseline model (Model RE, which is the MIT-B2 architecture) was initialized with random weights and trained directly on experimental data. All results shown here use the LOSO evaluation protocol described in the main text.

The results section of the main paper presents all the results for Dataset 1 (Bonnie *et al.*). In this supplementary information, all the results for Dataset 2 (Strantza *et al.*) are presented.

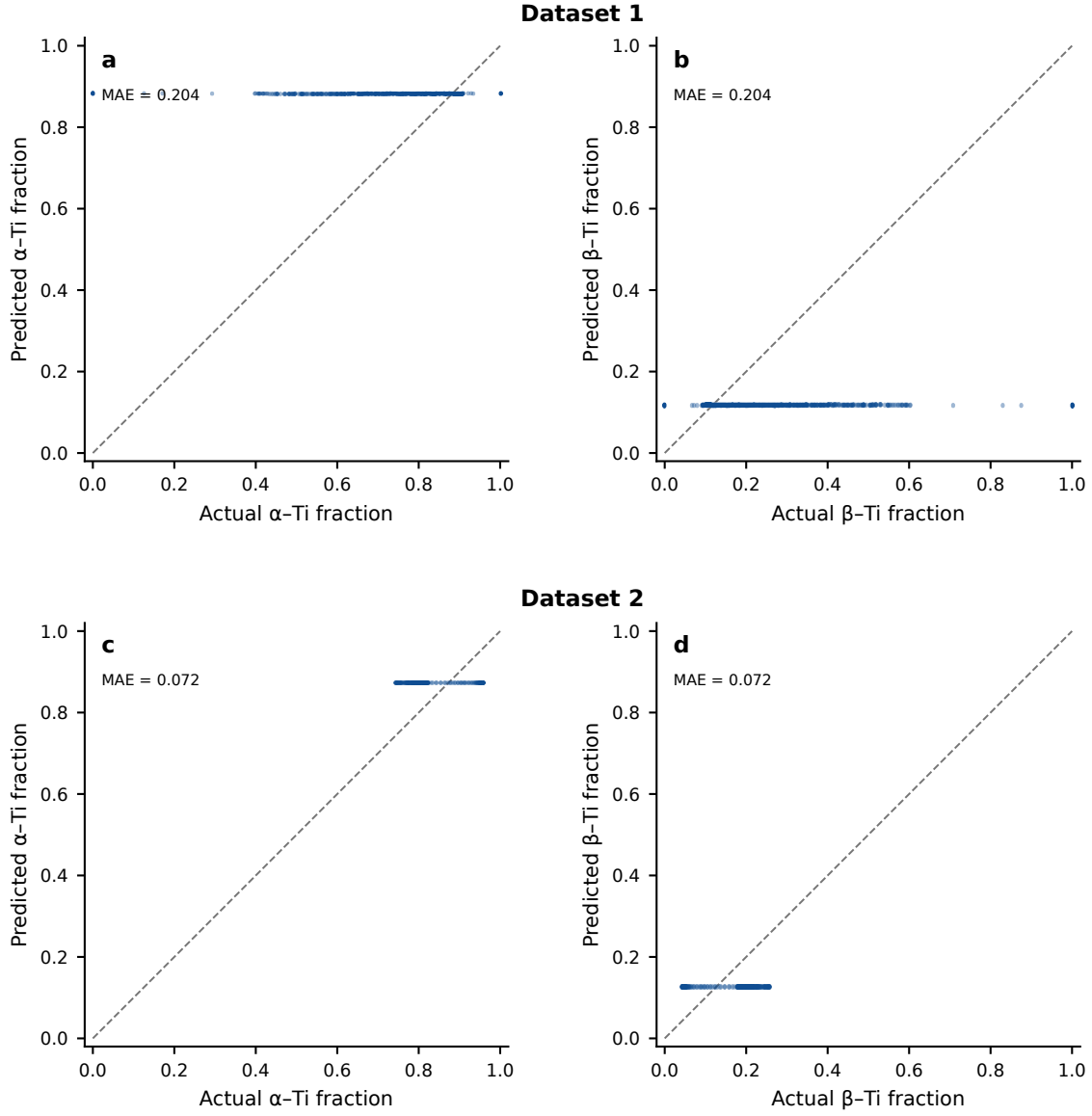


Figure S1: Performance of Model RE (random initialization, no pre-training) fine-tuned on $N_e = 250$ experimental diffractograms and evaluated on the LOSO test set. (a, b) Dataset 1, (c, d) Dataset 2. Predicted versus actual phase-volume fractions for α - and β -Ti.

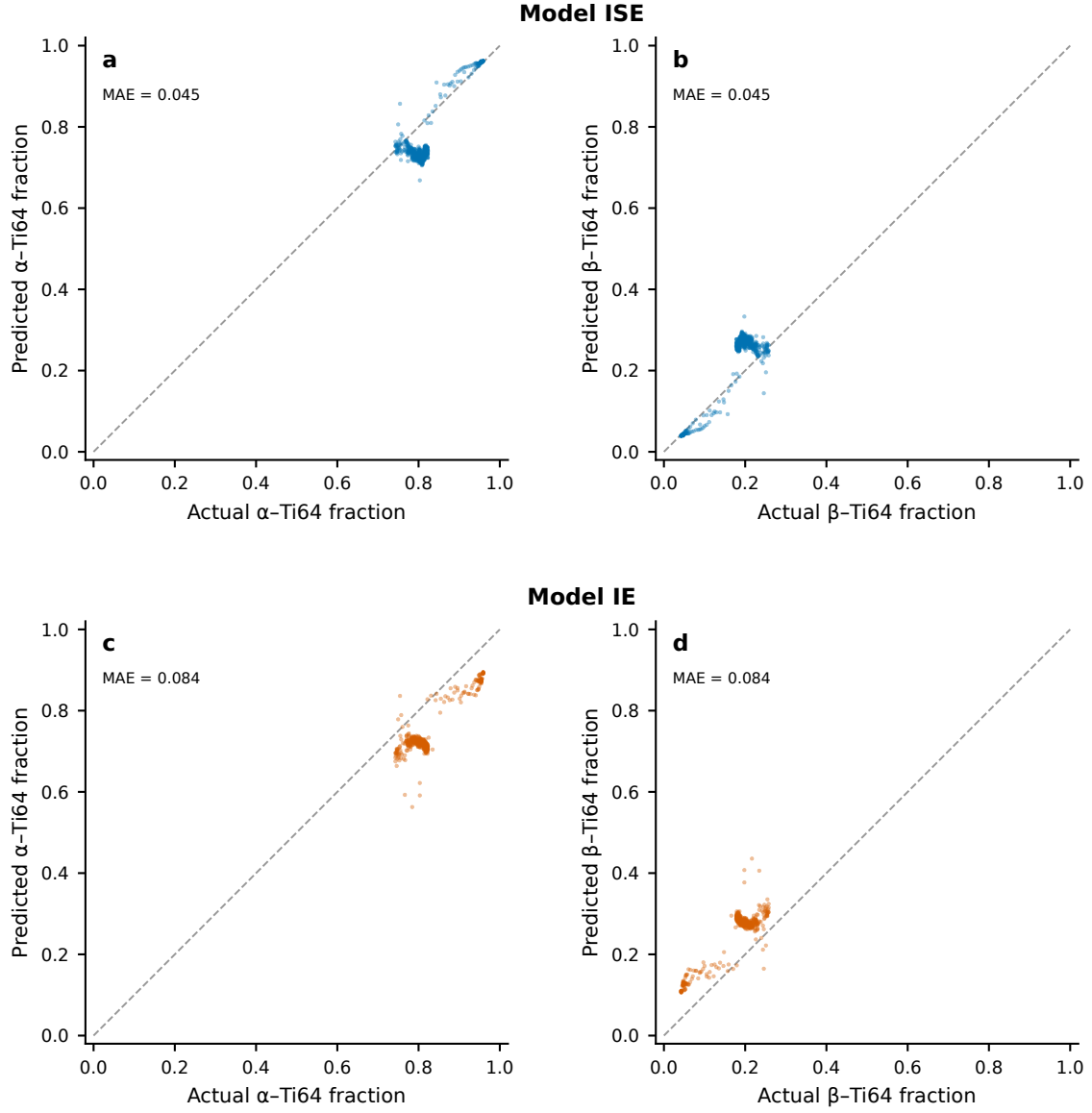


Figure S2: Performance of Models ISE and IE fine-tuned on $N_e = 250$ experimental diffractograms and evaluated on the LOSO test set of Dataset 2 (Strantza *et al.*). (a, b) Model ISE (MAE = 0.045), (c, d) Model IE (MAE = 0.084).

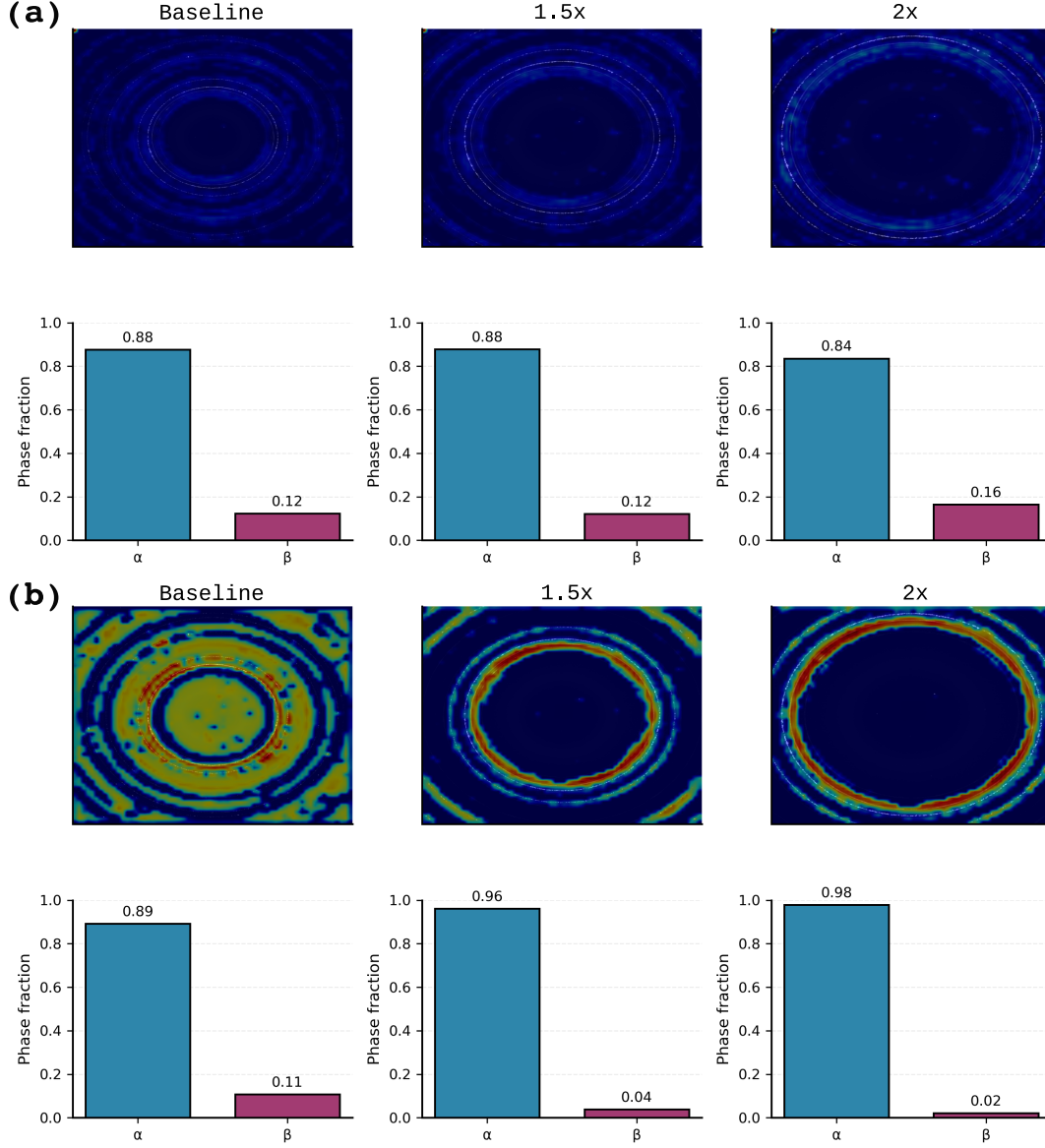


Figure S3: Geometrical stability test. Grad-CAM visualizations and corresponding phase-fraction predictions for Models IE and ISE under progressive geometric magnification of the input diffractogram. (a) Grad-CAM visualizations and phase-fraction predictions of Model IE under the zoom test. The diffractogram is progressively magnified ($1.0\times$, $1.5\times$, $2\times$), and the resulting Grad-CAM maps illustrate how the model's focus regions shift with geometric scaling. (b) Grad-CAM visualizations and phase-fraction predictions of Model ISE under the same zoom test conditions. The comparison highlights how geometric scaling alters the evidence each model relies on. In this case, as the pattern is magnified, most of the β -Ti rings disappear and the model correctly increases the α -Ti prediction. The test evaluates whether both models rely on scale-consistent features or whether their different optimization trajectories lead to distinct feature-use behaviors.

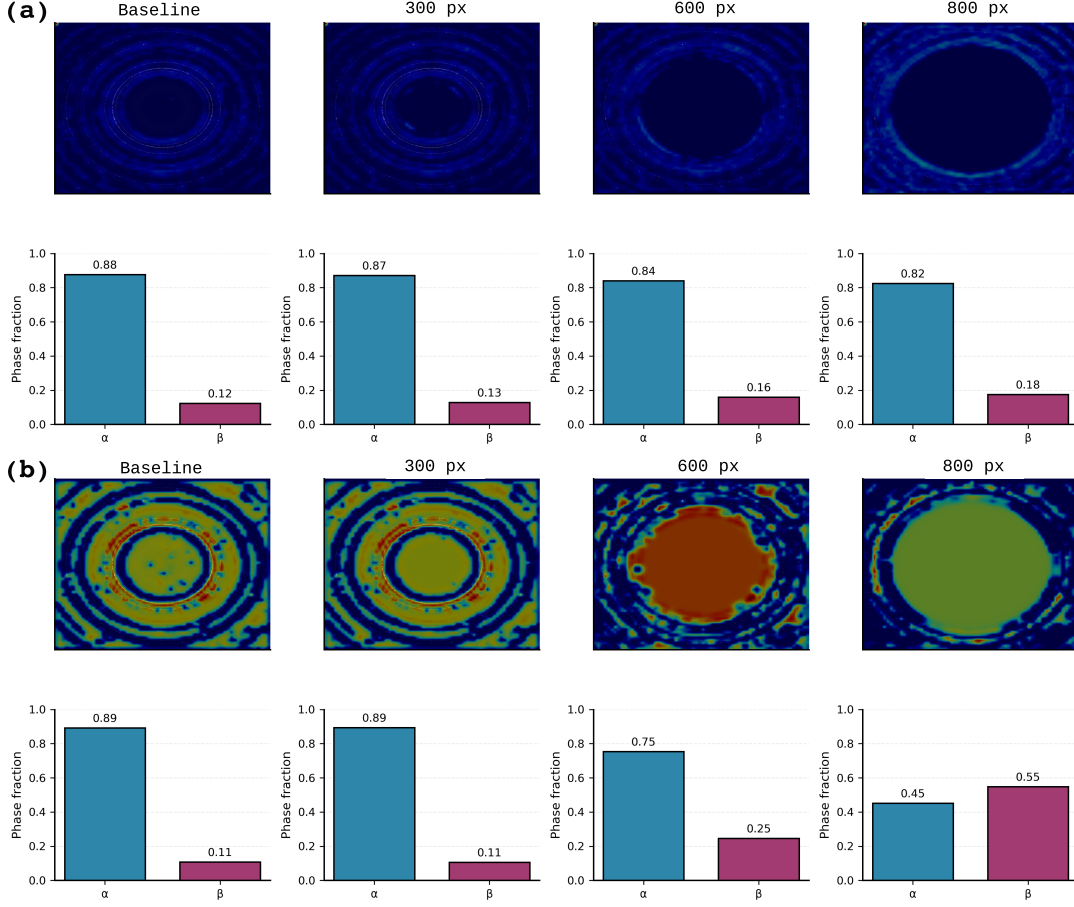


Figure S4: Physical causality test. Grad-CAM visualizations and corresponding phase-fraction predictions for Models IE and ISE under progressive radial masking of the input diffractogram. (a) Grad-CAM visualizations and phase-fraction predictions of Model IE under the physical causality test. The diffractogram is progressively masked radially (300, 600, and 800 pixels), and the resulting Grad-CAM maps illustrate how the model’s focus regions shift with masking of physical evidence. (b) Grad-CAM visualizations and phase-fraction predictions of Model ISE under the same physical causality test conditions. The comparison highlights how masking of physical evidence alters the evidence each model relies on. The test evaluates whether both models rely on physically-robust features or whether their different optimization trajectories lead to distinct feature-use behaviors.

For Dataset 2 (Strantza *et al.*), the geometric stability test requires a contextualized interpretation due to its larger experimental sample-to-detector distance. Because of the inherently narrower field of view, applying artificial magnification (zooming) quickly pushes the primary β -Ti diffraction rings outside the detector. Consequently, as the physical evidence for β -Ti is cropped out, a physically grounded model *should* correspondingly increase its predicted α -Ti phase fraction. Therefore, what might computationally appear as a geometric instability (a shift in predictions) is actually the desired, causally correct behavior. Figure S3 illustrates this phenomenon clearly: as magnification systematically excludes β -Ti peaks from the field of view, Model ISE appropriately adjusts its predictions to reflect the exclusively visible α -Ti evidence (Fig. ??).

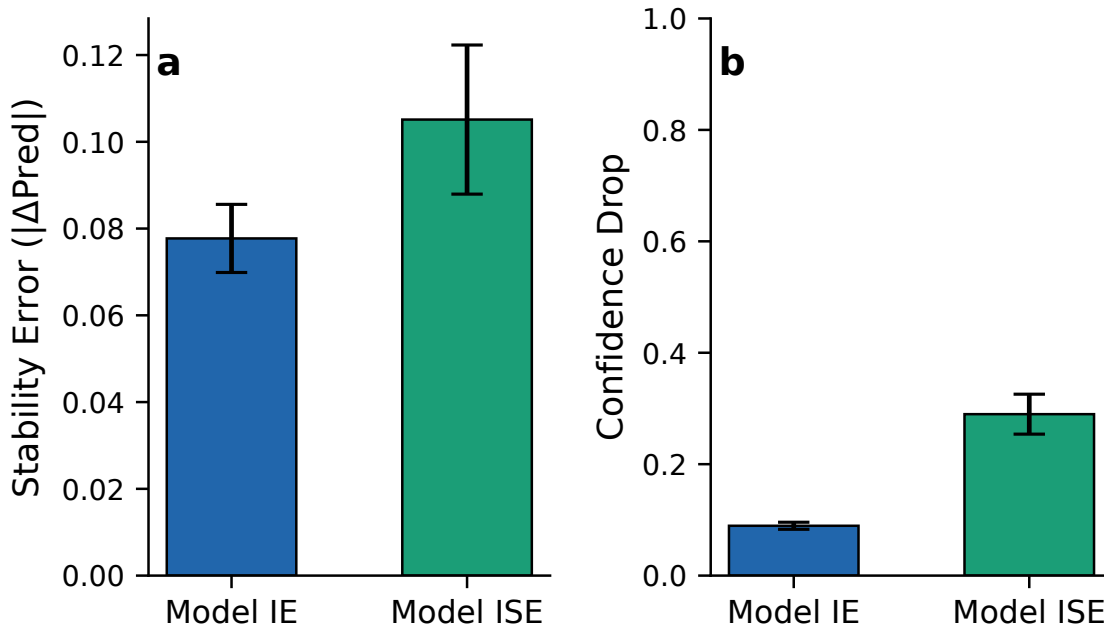


Figure S5: Quantitative comparison of Models IE versus ISE across 50 randomly selected diffraction patterns from Dataset 2. (A) Geometric stability measured via zoom perturbation test: higher values indicate better stability. (B) Physical causality awareness measured via radial mask test: higher values indicate stronger sensitivity to physically meaningful features. Error bars represent the standard errors.