The Shape Transfers, the Scale Does Not:

A Leakage-Free Benchmark and a Calibrated-Interval Deliverable for Curvature-to-Stress Prediction in Flax/PLA Hygromorph Biocomposites

Ethan Sheehan

University of Bristol , fc23871@bristol.ac.uk

July 24, 2026

Abstract

Active hygromorph biocomposites bend as their moisture content changes, and reading the internal surface stress that accompanies an observed shape is central to designing them. On 72 3D-printed continuous flax-fibre/poly(lactic acid) strips spanning 10 manufacturing batches we ask, under one leakage-free, batch-grouped protocol, what of the curvature-to-stress map can be recovered for an *unseen* batch. The constructive result is that the map *factorises* into a batch-invariant shape and a per-specimen scalar stiffness α ($\sigma \approx \alpha\, m(\Delta\kappa, s)$, pooled $R^2 = 0.92$ with an oracle scale), and that the *shape* genuinely transfers out of batch: from curvature alone, zero-measurement (Tier A) inputs reproduce the stress *pattern* of a held-out batch at a scale-free per-specimen weighted-r^2 of 0.82 (trained net) to 0.91 (deterministic binned master), well above the 0.5 bar—though the pattern is dominated by a shared monotone trend, so the specimen-specific increment over a shuffled-shape floor is modest. What does *not* transfer is the scale. Per-specimen effective stiffness spans ∼30–50× (between-batch ICC = 0.33–0.53); a variance decomposition using newly digitised per-specimen thickness and width shows this scatter is genuine material modulus, not a geometry artefact—the computable $t^3 w$ geometry term explains only ∼7% of $\mathrm{Var}(\log|\alpha|)$, and the stiffest batch survives geometry correction at 6.7×. No zero-measurement covariate recovers the scale: model class, conditioning time, specimen geometry, digitised thickness, microscope curvature, image texture, shape descriptors, domain-generalisation penalties (DANN, GroupDRO) and generative augmentation are all tested and excluded, whereas a single physical probe of the target specimen recovers it (per-specimen linear self-calibration $R^2 = 0.74$; one force–stroke reading ranks stiffness at Spearman 0.75 against 0.11 for zero-measurement). Because the between-batch scale spread is irreducible, raw batch count cannot buy point accuracy—a simulation calibrated to the data (passing 7/7 posterior-predictive checks) shows point R^2 plateauing near zero even at 30 batches while a calibrated interval reaches within 10% of its floor by ∼5 batches. The honest deliverable for an unseen batch is therefore a calibrated interval: a surrogate posterior-predictive band is the only method with ≥ 0.80 coverage on all 10 held-out batches (min 0.82 on the hardest) at 1.4× the width of pooled conformal, which collapses to 0.55 there. We release the protocol, frozen splits, and an experiment ledger as a reusable small-batch benchmark, and conclude that the binding constraint is stiffness *coverage*, not model capacity.

Keywords: hygromorph biocomposites; flax fibre; 4D printing; surface stress; local curvature; physics-informed machine learning; batch effects; domain generalisation; conformal prediction; small-data regression; leave-one-batch-out.

1 Introduction

Hygromorph biocomposites—bio-based laminates that change shape when their moisture content changes—are a route to passive, untethered actuators and morphing structures [1]. Continuous flax-fibre/poly(lactic acid) (flax/PLA) hygromorphs can be 3D-printed with programmable curvature [2], and their mechanical response depends strongly on moisture and on manufacturing quality [3, 4]. A practical design question is inverse to the usual shape-from-stimulus view: given the *shape* a specimen has taken (its curvature along the arc), what *surface stress* does the material carry? Answering this from data would let a designer read internal load off an observed profile.

To first order, stress and curvature are related by the Shape Actuation Modulus (SAM), which links a change in stress to a change in radius of curvature [5]; equivalently, by beam theory, the true surface stress is $\sigma \approx E\,(t/2)\,\Delta\kappa$, where $\Delta\kappa$ is the signed curvature change from the rest shape and $E\,t/2$ is an effective bending stiffness. This makes the problem attractive for a small, physically-structured model rather than a large generic one—a view reinforced by the material-neural-network programme of the same group, whose one trained network to date is a seven-node single-layer model grown by incremental data acquisition, albeit for a light-transmission actuator rather than a curvature–stress task [6, 7].

The dataset is characteristic of experimental composites work: 72 specimens, but because specimens within a manufacturing batch are near-replicates, the effective number of independent samples is closer to the *number of batches* (10) than to the specimen count. Such small, grouped datasets are easy to evaluate over-optimistically [8, 9], and—as a methodological caution we return to in §11—an earlier iteration of this very task reported a markedly rosier held-out score that was an artefact of test-set leakage. Our emphasis is therefore as much on *honest, exhaustive evaluation* as on any single model.

This paper is deliberately a systematic study rather than a single-method proposal, in the spirit of the domain-generalisation and recommender-system audits that found elaborate methods failing to beat simple baselines under fair evaluation [10, 11]. But it is framed around what *is* recoverable. Under one protocol we establish that the curvature-to-stress map separates cleanly into a transferable shape and an untransferable per-specimen scale, show which parts of the map cross a batch boundary and which do not, and—because the point prediction of an unseen batch is provably capped—deliver the calibrated interval that a designer can actually use.

Contributions.

1. **A leakage-free, batch-grouped benchmark** (§5) with frozen splits and a reproduction gate, offered as a reusable small-batch protocol. Every result below is comparable because it runs through the identical harness.
2. **The factorisation, and a positive out-of-batch result** (§7): the map is $\sigma \approx \alpha\, m(\Delta\kappa, s)$ (pooled $R^2 = 0.92$), and from curvature alone the *shape* of an unseen batch's stress field is recovered (scale-free per-specimen r^2 median 0.82–0.91); only the per-specimen scale does not transfer.
3. **A variance decomposition of the batch effect** (§6) built on newly digitised per-specimen thickness and width: the ∼30–50× stiffness spread is genuine material modulus (ICC = 0.33–0.53; the computable geometry term is only ∼7%), not a thickness artefact.
4. **An exhaustive exclusion** (§8): model class, every zero-measurement covariate (conditioning, geometry, digitised thickness, microscope curvature, image texture, shape descriptors),

domain-generalisation penalties, and generative augmentation are each tested and excluded; a single physical probe of the target specimen is the only thing that recovers the scale (§9).

5. **A calibrated-interval deliverable and a design rule** (§10): raw batch count cannot buy point accuracy (a data-calibrated simulation shows point R^2 plateauing near zero), so the correct output for an unseen batch is a calibrated interval—a surrogate posterior-predictive band covers all 10 held-out batches at ≥ 0.80—plus a request for one stiffness probe.

2 Related work

Flax-fibre hygromorph composites. The material system and its measurement conventions follow de Kergariou and colleagues: porosity and specimen-to-specimen variability in flax/PLA [4]; the near-exponential decay of stiffness with moisture [3]; the design and state of the art of 4D-printed hygromorphs [1]; and a programmable-curvature design space in which curvature is explicitly position-dependent along the specimen [2]. The SAM [5] provides the stress–curvature relation we exploit. The same programme has explored inverse design through evolutionary search over voxel finite-element models [12], and its material-neural-network line [6, 7] independently argues that data breadth, not network size, is the limiting resource—a conclusion the present study makes quantitative, noting that the one trained network in that line predicts light transmission in a wood/carbon-black shading machine, not stress in flax/PLA. Feedstock and process variability in natural-fibre printing is documented beyond this material system [13, 14]. To our knowledge, no prior work applies machine learning to the curvature-to-stress inverse problem, nor evaluates any hygromorph model under a grouped, batch-held-out protocol; this study is the first to do either.

Small-data materials ML. Reviews converge on injecting physical priors, preferring low-capacity models with strong classical baselines, grouping the resampling by the true unit of variation, and reporting uncertainty on the metric [8, 9]; ML on flax composites specifically remains data-limited [15, 16]. We follow this playbook, adding coordinate inputs in the spirit of Fourier features [17] but ultimately finding a $\sim$1.3k-parameter model sufficient, and we adopt the Δ-learning idea—predict a correction to a physical baseline—from quantum chemistry [18].

Batch effects and domain generalisation. A between-group scale shift that dominates prediction is the materials analogue of the *batch effect* long studied in genomics, where correction requires either replicated cross-batch controls or a measured covariate [19, 20]; batch-to-batch scatter in fused-deposition parts is itself documented [21]. Predicting an unseen batch is a domain-generalisation problem, and the fair-evaluation literature repeatedly finds that invariance penalties and adversarial schemes do not beat empirical risk minimisation [10, 11]; recovering a per-domain scale at test time is test-time adaptation [22, 23]. We baseline against Gaussian processes [24], gradient boosting [25, 26], hierarchical partial pooling [27, 28], physics-informed constraints [29, 30], conformal intervals [31, 32], and scaling-law extrapolation of the learning curve [33].

3 Problem setup and data

Specimens are rectangular 3D-printed continuous flax/PLA strips, moisture-conditioned until they actuate (bend) and then tensile-tested to failure while imaged. For each specimen and load state, the imaged centreline is reduced to coordinates (x, y) along an arc-length coordinate s, and the total surface stress σ is recorded along the arc. We resample every frame onto a common normalised-arc

grid of $N = 512$ points, $s \in [0, 1]$. The target is the field $\sigma(s)$ per load state; the inputs are the local curvature profiles of the rest shape $\kappa_{\text{nat}}(s)$ and the loaded shape $\kappa_{\text{app}}(s)$.

How stress and the per-specimen scale are computed. Stress is derived from the measured machine force and the deformed section geometry, $\sigma = \sigma_{\text{bend}} + \sigma_{\text{normal}}$ with $\sigma_{\text{bend}} = M\,(t/2)/(wt^3/12)$ and M the moment of the measured force; no material modulus is assumed, so a specimen's stress–curvature slope is a genuine measured stiffness. Two points about this provenance matter for the physics and are corrected here relative to earlier drafts. First, the section modulus in the stored pipeline was evaluated with a *fixed* thickness $t_0 = 1.1$ mm and width $w_0 = 10$ mm for every specimen, while the moment M carries the real geometry. Consequently the per-specimen stress–curvature slope

$$\alpha \;=\; \frac{\mathrm{d}\sigma}{\mathrm{d}\Delta\kappa} \;=\; \frac{6}{w_0 t_0^2}\,\frac{\mathrm{d}M}{\mathrm{d}\Delta\kappa} \;=\; \frac{E\,w\,t^3}{2\,w_0 t_0^2} \tag{1}$$

is proportional to the specimen's true bending stiffness $E\,w\,t^3$, in which *thickness enters cubically*—not the surface-strain form $E\,t/2$ that describes the true surface stress with a consistent thickness. The stored σ and the true σ coincide only when $t = t_0$; where they differ, α is best read as the real $E\,I$ divided by a frozen section modulus. This is a factual correction, not a change of interpretation: α remains a per-specimen measured stiffness, and we show in §6 that the discarded geometry contributes little to its scatter. Second, we have since digitised per-specimen thickness t and width w from the side-view images in each specimen folder (silhouette half-width at the medial axis, calibrated by the gauge length). The specimens are nominally identical in section; the digitised thickness is correspondingly tight, 1.16 ± 0.09 mm, and because silhouette segmentation and residual image-scale differences inflate this figure, its CV $\approx 8\%$ is an *upper bound* on the true thickness variation. This lets us bound the geometry term conservatively rather than assume it away.

We resample and treat α throughout as the per-specimen weighted least-squares slope of σ on $\Delta\kappa$ (a Tier C oracle in the sense of §5). The 72 usable specimens were manufactured in 10 batches (Table 1); sizes $\{2, 3, 6, 7, 10, 10, 10, 8, 9, 7\}$. Because specimens within a batch share layup and conditioning they are statistical near-replicates: the effective sample size is $\sim$10, not 72. Global natural curvature ranges 0.008–0.041 mm^{-1}, specimen length 52.7–62.4 mm, load states per specimen 35–301, and peak stress 0.26–10.4 MPa; the high-stress tail belongs almost entirely to one batch (§6). Table 2 inventories the covariates and their status—the recurring obstacle is that the only recorded per-batch covariate, conditioning time, is nearly collinear with batch identity and missing for the batch that matters most.

4 Curvature representation

A single scalar $\kappa = 1/R$ discards real position-dependent variation (a diagnostic finds the local profile varies by a median $\sim$70% of its mean, reproducibly across load states), so we estimate a *local* signed profile $\kappa(s)$. Because this is an ill-posed differentiation of a noisy centreline, we treat estimator choice empirically, scoring seven estimators on all 72 specimens against three physical gates: consistency with the global circle fit, shape-reconstruction fidelity, and cross-load reproducibility (Table 3). A signed generalised-cross-validation smoothing spline is selected ($r = 0.99$ reproducibility, essentially zero roughness, consistency 0.91) while the intuitive sliding-window local circle fit is *rejected* as ill-conditioned. The single most consequential curvature error in the earlier pipeline was a dropped sign (an absolute value), which destroys the signed $\Delta\kappa$ that SAM requires—not the

Table 1: The 10 manufacturing batches. Conditioning time is per batch (from the naming convention; batch 10 undocumented). |SAM slope| = $|\mathrm{d}\sigma/\mathrm{d}\Delta\kappa|$ is the per-batch median; the batch medians span ∼53×, and the per-specimen range on reliable fits ($r^2 \geq 0.6$) is ∼30× (§6).

Batch	Spec. (n)	Cond. (h)	Med. $\sigma_{\max}$ (MPa)	|SAM slope| (MPa mm)	Med. $\lvert\Delta\kappa\rvert_{\max}$ (mm^{-1})
1	2	24	0.88	9.5	0.026
2	3	48	3.79	305.5	0.016
3	6	43	1.98	67.7	0.032
4	7	28	1.11	64.8	0.020
5	10	43	2.10	81.1	0.028
6	10	24	1.99	73.5	0.026
7	10	28	1.98	104.0	0.016
8	8	43	3.45	165.4	0.021
9	9	25	0.58	25.5	0.020
10	7	—	**7.29**	**505.6**	0.015

Table 2: Covariate inventory. Only conditioning time is a recorded per-batch covariate, and it is near-collinear with batch. Digitised thickness/width, microscope curvature and image texture are zero-measurement (Tier A) quantities of the target; the force–stroke initial stiffness is a physical probe of the target specimen (Tier B, §5); true microstructure (porosity, fibre volume fraction) was not measured in this dataset.

Covariate	Status	Use
Local curvature fields $\kappa_{\mathrm{nat}}, \kappa_{\mathrm{app}}$	measured	model input (Tier A)
Arc position s, specimen length	measured	model input (Tier A)
Natural-curvature statistics	derived	scale predictor (Tier A)
Digitised thickness t, width w	digitised (side view)	scale predictor (Tier A)
Microscope curvature κ_{img}, image texture	digitised (62–69/72)	scale predictor (Tier A)
Conditioning time (h)	inferred, batch-collinear	scale predictor (Tier A)
Force–stroke initial stiffness	measurable (target's own test)	scale probe (Tier B)
Microstructure (porosity, V_f)	*not measured*	—

smoothing method. We adopt signed spline $\kappa(s)$ and carry a per-point uncertainty equal to the inter-estimator spread.

5 Evaluation protocol (the benchmark)

Because specimens within a batch are near-replicates, the resampling unit is the *batch*. We report two generalisation questions, plus a self-calibration protocol, all on *frozen, versioned splits* loaded by every experiment:

- **Within-batch** (grouped): hold out a fixed fraction of specimens from every batch; all batches, hence all α, are represented in training. New specimen of a *known* batch.
- **Leave-one-batch-out (LOBO)**: hold out an entire batch (10 folds). *Unseen* batch; the held-out scale is unknown and is set to the mean of the trained batches unless calibrated.
- **Self-calibration**: for an unseen specimen, calibrate its scale on its own first k (lowest-load) frames and predict frames $> k$—honest load-extrapolation.

Table 3: Local-curvature estimators, median over 72 specimens. Consistency = arc-mean $|\kappa(s)|$ over the global circle-fit $1/R$ (ideal 1); reproducibility = cross-load shape correlation; roughness = scale-normalised |2nd difference|. The signed spline is selected; the local circle fit is rejected.

Estimator	Recon. err. (% arc)	Consistency	Reproducibility	Roughness
Signed spline (selected)	8.03	**0.91**	**0.99**	**0.00**
Savitzky–Golay (28% arc)	7.98	0.81	0.97	0.00
Savitzky–Golay (16% arc)	8.01	0.90	0.95	0.01
Savitzky–Golay (8% arc)	8.03	1.14	0.82	0.05
Local circle fit (window 24)	8.76	3.57	0.37	0.97
Local circle fit (window 16)	11.68	5.61	0.27	1.27
Menger three-point	8.02	9.77	0.06	2.22

No held-out information is used for standardisation, early stopping, or model selection (early stopping uses an inner split of the training batches only). Metrics are pooled over held-out points (pooled RMSE and R^2) and, for the unseen-batch question, also per-batch-averaged (macro R^2, each batch counting once); uncertainty is a batch-level bootstrap 95% CI, and every neural result is reported as a mean ± standard deviation over 5 seeds, because at this data scale seed variance is itself substantial.

Feature-honesty tiers. Since σ is derived from the measured force, any feature containing force at an *evaluated* frame is circular. We therefore label every experiment: **Tier A** uses only zero-measurement quantities of the target (curvature fields, geometry, digitised thickness, microscope curvature, image texture, conditioning); **Tier B** uses one physical probe of the target specimen (its own early frames, or a force–stroke stiffness reading from a low-load window); **Tier C** (circular) is reported only as an oracle upper bound. *The headline generalisation claim is judged on Tier A only*; Tier B results are the practical upper bound.

Reproducibility. The harness re-derives the frozen reference model's numbers exactly for the linear baselines and within tolerance for the neural models (a reproduction gate), and every table cell traces to a row in a machine-readable experiment ledger. Code, splits, and ledger accompany the paper.

6 The batch effect: a variance decomposition

Fitting σ against $\Delta\kappa$ per batch (the SAM slope) gives batch medians from 9.5 to 505.6 MPa mm, a ∼53× range (Table 1, Fig. 1a); the extreme low end is batch 1 ($n = 2$, fit $r^2 \approx 0.12$), so on reliable per-specimen fits ($r^2 \geq 0.6$) the honest range is ∼30×. Batch 10 (specimens 81c–89c), whose peak stresses run 3–5× the median, has the steepest slope (≈ 6.9× the other-batch median) while its curvature change is if anything *smaller* than average: its high stress is a stiffer material response, not larger deformation.

Is the scale a geometry artefact? Because α scales as $E\,w\,t^3$ (Eq. 1) and thickness was frozen in the stored pipeline, one must ask whether the spread is real modulus or discarded geometry amplified by the cube. The digitised thickness answers it. Decomposing $\log|\alpha|$ over the identity $\mathrm{Var}(\log|\alpha|) = \mathrm{Var}(\log E) + \mathrm{Var}(\log t^3 w) + 2\,\mathrm{Cov}$, the computable $t^3 w$ geometry term carries only ∼7% of the total $\mathrm{Var}(\log|\alpha|) = 1.62$ (free-fitting $\log(t^3 w)$ explains $R^2 = 0.018$ of $\log|\alpha|$); the

residual material modulus carries the rest. The digitised thickness is far too uniform—its CV $\approx 8\%$ is itself an upper bound inflated by segmentation and image-scale noise, and even taken at face value spans at most a $\sim$2.5$\times$ range in t^3—to manufacture a $\sim$30–50$\times$ scale range. Crucially, batch 10's thickness is normal (1.157 mm vs. the dataset median 1.156 mm), and its stiffness *survives* geometry correction: its α is 7.3$\times$ the other batches, and after removing $t^3 w$ the corrected modulus E is still 6.7$\times$. Recomputing σ with the measured t, w shrinks $\mathrm{Var}(\log|\alpha|)$ by only 2.5% and leaves the master-curve collapse essentially unchanged (§7). A second geometric confound—per-specimen image scale, since a wrong pixel-to-millimetre factor f would inflate α as f^2—is equally ruled out: the recovered gauge length is constant to CV = 1.9%, $\log|\alpha|$ is uncorrelated with it (between-batch $r = -0.26$, within-batch $r = -0.11$, both non-significant and of the *wrong sign* for an f^2 inflation), and dividing α by f^2 *widens* the spread rather than shrinking it. Neither the discarded thickness nor the image scale explains the range: the scale is genuine material stochasticity, now *proven* rather than assumed (Fig. 1b).

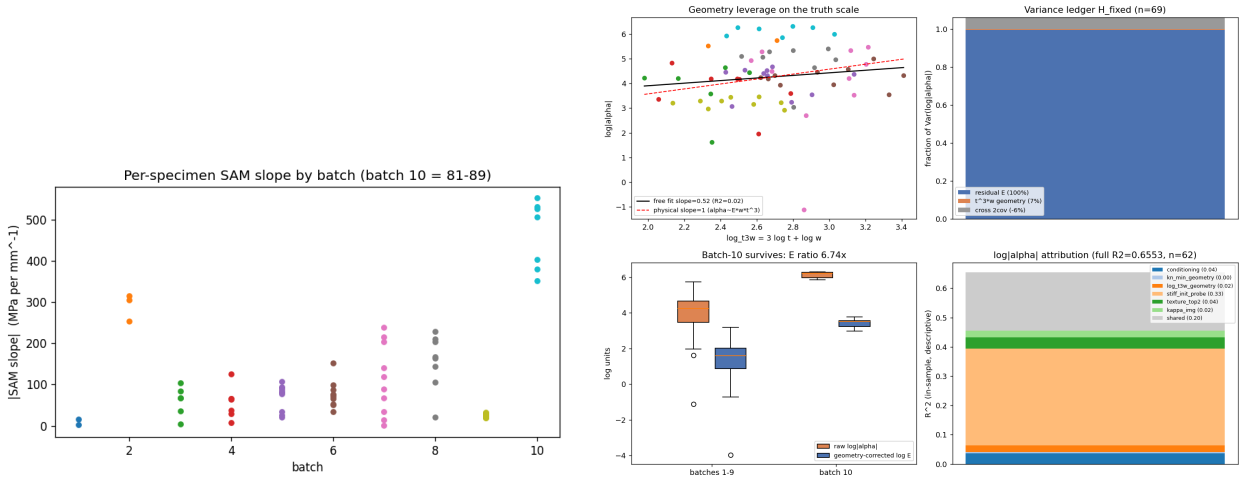
Decomposing $\log|\alpha|$ into between- and within-batch components gives an intraclass correlation of ICC = 0.33 over the 9 documented batches (0.53 including batch 10)—so roughly a third to a half of the stiffness variance is between batches, and the majority is *within* batch. This is the quantitative statement of why the problem is hard: the scale that must be predicted is a per-specimen quantity with a large between-group component, and (as Table 2 shows) every covariate that might index it is confounded with the batch label.

Disclosure: batch 10 is a different acquisition. Batch 10 is the linchpin of the scale story, and honesty requires flagging that it differs from the other 65 specimens on three uncontrolled axes at once, not only in stiffness: it was imaged on a different camera (2848×2848 vs. 2464×2056), with a different pixel-to-mm calibration (≈ 0.027 vs. an identical 0.0427 copied across the fixed-rig specimens), and its conditioning time is undocumented. The calibration self-validates (every batch, including batch 10, reproduces a $\sim$60 mm gauge length, CV 1.9%), so this is a genuinely stiffer material on a correctly-calibrated but different rig—but the confound is disclosed and batch 10 is treated as the hardest, not the representative, held-out batch throughout.

A validated generative surrogate. To reason about how many batches would help and what interval to report (§10), we fit a two-level generative model of the scale: batch means $\theta_b \sim \mathcal{N}(\mu_0, \tau^2)$ and specimen scales $x_i \sim \mathcal{N}(\theta_b, \omega^2)$ on $\log|\alpha|$, with the rank-1 shape $m(\Delta\kappa, s)$ and a heteroscedastic residual model. Fitted, it recovers $\mu_0 = 4.19$, $\tau = 1.04$, $\omega = 0.97$ (ICC = 0.53) and reproduces α to machine precision. It passes all 7/7 posterior-predictive checks—ICC, τ, batch-median $|\alpha|$ range, global-linear R^2, the master-curve collapse R^2, the spread of per-batch log-slopes, and the conditioning correlation are all inside their simulated $[0.05, 0.95]$ predictive intervals (Fig. 2)—so it is a legitimate stand-in for the population when we simulate augmentation (§8), batch power and intervals (§10).

7 The factorisation: the shape transfers

The central positive result is structural. If we grant each specimen its own scalar α, a single batch-invariant shape function reproduces the whole dataset: $\sigma \approx \alpha\, m(\Delta\kappa, s)$ reaches pooled $R^2 = 0.92$ (0.84 without the arc term s), versus 0.85 for unconstrained per-specimen lines and only 0.17 for one global line (Fig. 3). A per-*batch* scale reaches only 0.71, confirming that the scale lives at the specimen, not the batch, level. So the curvature-to-stress map genuinely factorises into a transferable shape and a one-number stiffness, and the entire generalisation problem reduces to



(a) Per-specimen SAM slope by batch

(b) Variance attribution of $\log|\alpha|$

Figure 1: The batch effect is real material stiffness. (a) Per-specimen SAM slope $|\mathrm{d}\sigma/\mathrm{d}\Delta\kappa|$ by batch; batch 10 is $\sim$6.9$\times$ stiffer than the median batch. (b) Decomposition of $\mathrm{Var}(\log|\alpha|)$: the computable t^3w geometry term (from digitised thickness/width) is only $\sim$7%; the rest is residual material modulus. Batch 10 survives geometry correction at 6.7$\times$.

estimating that one number for a new specimen. The collapse is insensitive to binning (R^2 in 0.905–0.921 across $\Delta\kappa \times s$ bin counts) and, per §6, is unchanged when σ is recomputed with the measured per-specimen t, w.

The shape transfers out of batch. The factorisation would be hollow if the shape itself did not cross a batch boundary. It does. Under LOBO—training the shape on nine batches and predicting the tenth—we score the recovered stress *pattern* with a scale-free per-specimen weighted-Pearson r^2 (invariant to the missing scale) and an optimal-scale R^2. A trained scale-free shape net reaches a median per-specimen shape-r^2 of 0.82 (macro 0.75), and a deterministic binned master curve reaches 0.91 (macro 0.81); both are far above the 0.5 bar (Table 4, Fig. 4). This is the paper's first legitimate positive out-of-batch result: from curvature alone, one can predict what the stress field of an unseen batch will *look like*. Two honest caveats. First, the deterministic binned master out-predicts the trained net, so the useful object is the factorised master curve, not a learned network. Second, the field is dominated by a shared monotone $\Delta\kappa$ trend: the matched binned r^2 (0.971) exceeds a shuffled-shape floor (0.944), confirming a real specimen-specific component, but that shared trend is itself a high baseline, so the specimen-specific increment is modest. What does *not* transfer is the absolute scale: predicting stress in physical units for an unseen batch gives pooled $R^2 = 0.02$ and macro -1.73. The shape transfers; the scale does not.

The factorisation is visible directly in the predicted stress fields (Fig. 5) and in slope recovery. Fitting the model's predicted σ–$\Delta\kappa$ slope per specimen and comparing to the measured one, the within-batch, scale-known model recovers the per-specimen stiffness to a median factor of 1.9$\times$ (sign agreement 100%), whereas the unseen-batch model is off by 2.9$\times$ overall and 6.7$\times$ on the stiff batch 10—the error is almost entirely in the missing scale, not the shape.

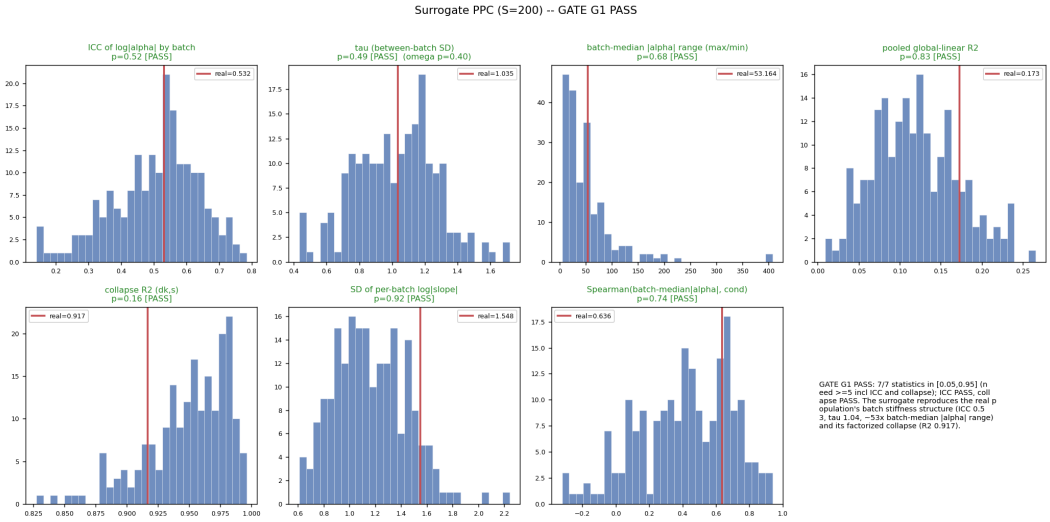


Figure 2: Posterior-predictive checks for the generative surrogate. Seven population statistics (real values marked) all fall inside the surrogate's simulated predictive intervals (7/7 pass, including ICC and the master-curve collapse), legitimising the simulation-based analyses of §8 and §10.

Table 4: The shape transfers out of batch; the scale does not (LOBO, mean over 5 seeds). The scale-free per-specimen weighted r^2 is invariant to the missing scale; the absolute columns predict stress in physical units. The binned master curve is the strongest, and the specimen-specific component sits above but near the shuffled-shape floor. Bar for a meaningful shape: $r^2 \geq 0.5$.

	Scale-free r^2		Real vs. shuffle	Absolute R^2	
Predictor (Tier A)	median	macro	(binned r^2)	pooled	macro
Binned master curve $m(\Delta\kappa, s)$	**0.91**	**0.81**	0.985 / 0.961	+0.06	−1.47
Trained shape net	0.82	0.75	0.971 / 0.944	+0.02	−1.73
Physics net (per-batch scale)	0.76	0.61	0.929 / 0.930	−0.28	−3.44

8 What does not close the gap

Table 5 is the roadmap: every method on the identical folds and metric. The shape transfers (§7); everything in this section is an attempt to recover the missing *scale* from something cheaper than a measurement, and every one fails. We report both pooled and per-batch-averaged (macro) R^2 for the unseen case, because pooled R^2 is dominated by the high-point-count central-stiffness batches and flatters methods that predict a middling response.

Model class and capacity (no). Within a known batch, every reasonable model works and they cluster: the curvature-only net ($R^2 = 0.50 \pm 0.22$), gradient boosting (0.65), the additive physics net (0.54 ± 0.15), the per-batch-linear baseline (0.56) and the scale-separable net (0.51 ± 0.05) all sit in a 0.5–0.65 band; a degree-2 ridge (0.20) and the global-linear baseline (0.12) trail. This is close to a measurement-noise ceiling: propagating the per-point curvature uncertainty through $\sigma \approx \alpha\Delta\kappa$ gives a winsorised ceiling of $R^2 \approx 0.65$ on the seven central, compliant batches. On an *unseen* batch the picture inverts. On pooled R^2 the only positive methods are the trivial global-linear SAM baseline (0.09) and a low-capacity Gaussian process (0.17), which reaches that score by fitting the central batches while failing the tails; on the honest macro metric *every* method is strongly negative, from

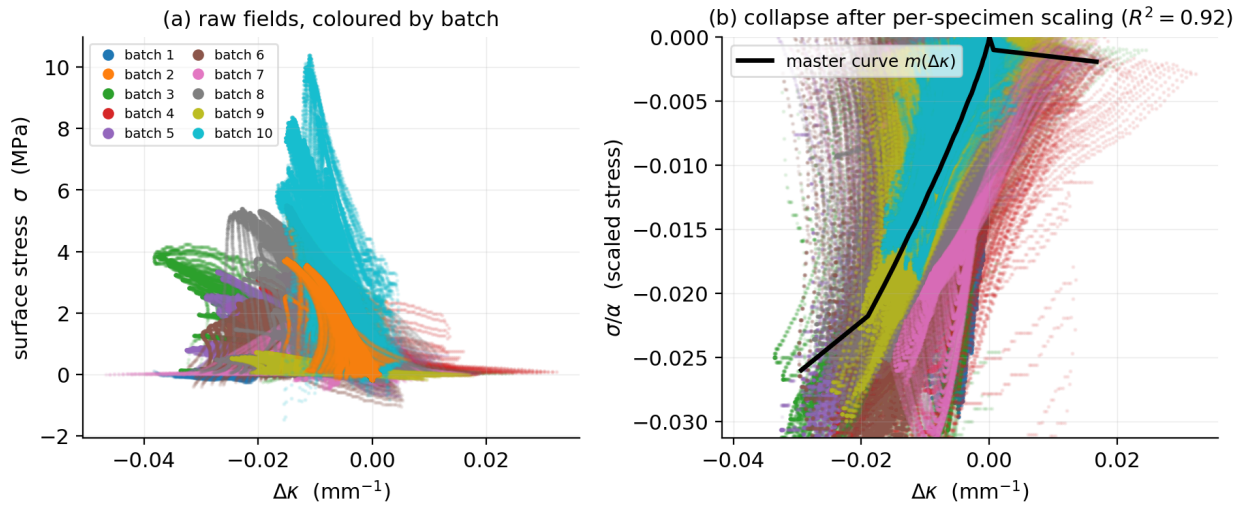


Figure 3: Master-curve collapse. (a) Raw stress fields separate by batch, batch 10 reaching the highest stress. (b) Scaling each specimen's stress by its own scalar α collapses the batches onto a single invariant shape $m(\Delta\kappa)$ (pooled $R^2 = 0.92$ with the arc term s; degenerate near-flat specimens excluded). The shape is batch-invariant; only the scale α varies.

-2.4 (curvature-only net, boosting) to -4.6 (GP) to -7.8 (per-batch-linear). **Capacity is not the lever, and the wall is metric-invariant.**

Zero-measurement covariates (no). The clean fix would be to predict α from something known without touching the specimen. Conditioning time carries a real but weak signal (permutation-test $\rho = 0.72$, $p = 0.036$; mixed-effects slope $+0.058 \pm 0.026$ log-units/h), but at $n = 9$ batches only $|\rho| \geq 0.68$ is detectable, batches sharing a conditioning value still differ several-fold, and injecting a conditioning-predicted scale does not beat the cold start (LOBO $R^2 = -0.20$ vs. -0.28). We then screened every newly digitised Tier A family for out-of-batch prediction of $\log|\alpha|$ (Table 6): specimen geometry and natural-curvature statistics, digitised thickness/width, scalar microscope curvature, a 20-dimensional image-texture set, and shape descriptors with a fold-safe PCA of the curvature field. *All are negative out-of-batch* (best, digitised dimensions, $R^2 = -0.08$; target-shuffle permutation $p = 0.69$), and their within-batch-centred variants are near zero—so whatever weak association exists lives in the illumination/session-confounded batch channel, not inside a batch. The only strong predictor is the Tier B force–stroke probe (§9). **No zero-measurement covariate makes an unseen batch predictable.**

Changing the curvature input (no). A natural reviewer question is whether a pre-test, force-free microscope image of the specimen would work better than the test-frame curvature. The two curvature channels agree well as scalars (Pearson $r = 0.87$ over the 51 covered specimens), but they are not interchangeable inside the factorisation: replacing the spatial test-frame curvature *field* with a scalar microscope κ_{img} collapses the master curve from pooled $R^2 = 0.79$ to -0.32 (macro $0.75 \to -0.67$) on the 62 covered specimens (Fig. 6). The position-dependent field is load-bearing; a scalar cannot substitute for it.

Domain-generalisation penalties (no). Treating the batch as a domain invites the DG toolbox. A gradient-reversal domain-adversarial network (DANN) trained on the batch-agnostic trunk

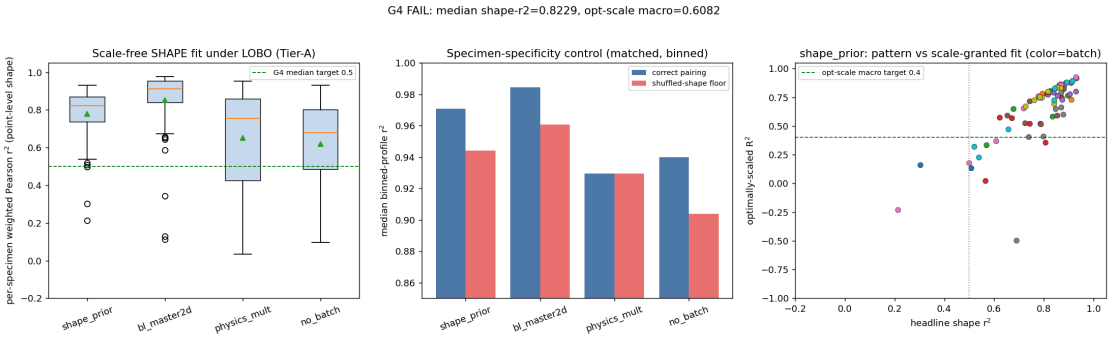
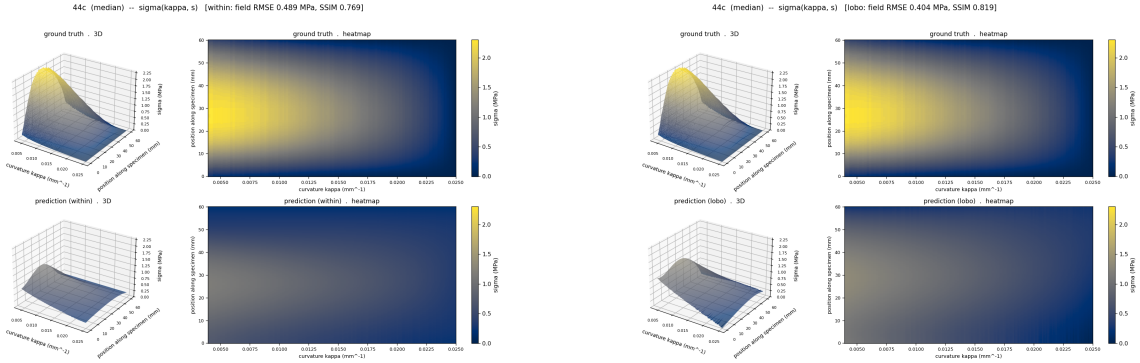


Figure 4: Scale-free out-of-batch shape recovery. Per-specimen weighted-Pearson r^2 of the predicted vs. measured stress pattern under LOBO, by predictor and by batch; the deterministic binned master (median 0.91) leads the trained net (median 0.82), both above the 0.5 bar. The absolute scale (not shown) does not transfer.



(a) Known batch (scale seen) **(b)** Unseen batch (scale unknown)

Figure 5: Measured vs. predicted surface-stress field $\sigma(\kappa, s)$ for a median specimen (44c). The predicted *shape* of the field is right in both cases; the unseen-batch prediction differs mainly by an overall scale. Data-space SSIM 0.77 (known) and 0.82 (unseen); field RMSE 0.49 and 0.40 MPa. For central-scale batches the batch-mean scale is close, so the unseen field is not always worse—the collapse appears only for scale-extreme batches.

does not beat the global-linear baseline on unseen batches (LOBO pooled $R^2 = -0.28$, macro -2.30), and a GroupDRO variant is no better (pooled -0.16, macro -1.46)—exactly as the fair-evaluation literature predicts [10, 11]. The reason is structural: the between-batch shift is, to first order, a single multiplicative stiffness (§7) that the shape function has already factored out, so an invariance penalty has nothing to remove that a one-number calibration does not already provide. DG is the wrong tool for a scale-shift domain problem; test-time calibration [22, 23] is the right one, and even it needs a measurement.

Generative augmentation (no—and it hurts). If the missing scale is a wide, sparsely-sampled distribution, one might hope to fill it in with synthetic batches drawn from the validated surrogate (Fig. 2). We trained on the real batches plus 10 synthetic ones and compared unseen-batch transfer to the real-only baseline, paired per seed. Augmentation *hurts* in every seed: at the fitted between-batch spread, macro-R^2 falls by 4.7 ± 3.6 (0/5 seeds positive; pooled -0.9), and widening the synthetic spread $1.5\times$ makes it worse (-8.7 ± 4.5; pooled -1.2), a clean dose-response (Fig. 7). The batch-10 conformal band is essentially unchanged. You cannot manufacture the

Table 5: Master comparison (leakage-free; ± over 5 seeds where applicable). Within-batch = new specimen of a known batch; unseen = LOBO. *On the macro metric every method is strongly negative*; the positive pooled scores (GP, global-linear) come from the high-point-count central-stiffness batches, not from generalising to the tails. Domain-generalisation methods (DANN, GroupDRO) and generative augmentation are no exception.

		Within-batch	Unseen-batch (LOBO)	
Method	Tier	R^2 (pooled)	R^2 (pooled)	R^2 (macro)
Curvature-only net	A	+0.499±0.224	-0.165±0.115	-2.417±0.535
Physics net (per-batch scale)	A	+0.512±0.045	-0.278±0.149	-3.435±0.296
Physics net (additive residual)	A	+0.538±0.154	-0.221±0.099	-2.430±0.472
Gradient boosting	A	+0.654	-0.072	-2.486
Gaussian process	A	+0.204	+0.165	-4.625
Ridge (poly-2)	A	+0.197	+0.029	-6.970
Global linear (SAM)	A	+0.118±0.000	+0.094±0.000	-3.775±0.000
Per-batch linear	A	+0.559±0.000	-0.055±0.000	-7.808±0.000
Predict mean	A	-0.001±0.000	-0.061±0.000	-3.400±0.000
+ conditioning-predicted scale (Tier A)	A	–	-0.202±0.119	-2.612±0.205
+ geometry-predicted scale (Tier A)	A	–	-0.155±0.106	-3.073±0.581
+ force-stroke probe scale (Tier B)	B	–	-0.708±0.636	-2.533±0.277
+ 3-shot batch calibration (Tier B)	A	–	-0.009±0.040	-2.656±0.215
Domain-adversarial (DANN)	A	+0.288	-0.278	-2.303
GroupDRO	A	+0.670	-0.159	-1.460
+ synthetic-batch augmentation (Tier A)	A	–	-1.169±1.510	-8.127±3.318

missing scale from a model of the scale you already have.

9 What one measurement recovers

If the scale cannot be predicted, it can be *measured*, and a single cheap probe of the target specimen is enough. Two results make this concrete.

Self-calibration from the specimen's own early frames. Few-shot batch calibration—fit the unseen batch's scale from k of its specimens—barely helps (3-shot $R^2 \approx 0$), because between-batch differences are not a clean global rescaling. Self-calibration on a specimen's *own* low-load frames is far more effective, with an instructive twist (Fig. 8a): calibrating the neural model's scalar *hurts* ($R^2 < 0$), whereas a plain per-specimen *linear* fit on the same early frames, extrapolated to higher loads, climbs steadily to $R^2 = 0.17, 0.42, 0.67, 0.74$ at $k = 3, 5, 10, 20$ frames. Two calibration frames are not enough—the k=1 rest frame has $\Delta\kappa \equiv 0$ and the slope is undefined, and recovering the full-frame oracle scale needs ∼10–20 loaded frames (Spearman of α_k vs. α_{full} rises $0.54 \to 0.93$ from k=5 to k=20)—but once observed, the transferable component is the *linearity* of σ in $\Delta\kappa$, not the learned shape.

One probe orders specimens by stiffness. A designer often needs only to rank specimens, not to predict stress exactly. Here the tier contrast is sharp (Table 7, Fig. 8b). Zero-measurement Tier A features *cannot* rank specimens by stiffness out of batch (Spearman $\rho = 0.11$, permutation $p = 0.16$, between-batch C-index 0.54, tertile accuracy 0.35 ≈ chance). A single force–stroke probe (Tier B) ranks them *strongly* ($\rho = 0.75$, $p = 0.001$, between-batch C-index 0.80, tertile accuracy

Table 6: Out-of-batch prediction of $\log|\alpha|$ (LOBO, pooled R^2; best of ridge/RF). Every zero-measurement (Tier A) family is tested and excluded; only the Tier B probe (a direct stiffness measurement of the target) carries signal.

Feature family	Tier	LOBO R^2
Conditioning + κ-statistics (reference)	A	-0.29
Digitised thickness/width (t, t^3w)	A	-0.08
Microscope curvature κ_{img} (scalar, 62 covered)	A	-0.14
Image texture (20-d, 62 covered)	A	-0.34
Shape descriptors + fold-safe PCA(6) of κ-field	A	-0.47
Force–stroke probe (target's own test)	B	$+0.40$

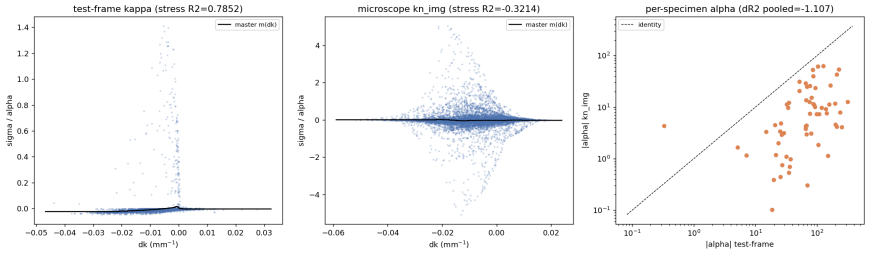


Figure 6: The spatial curvature field is load-bearing. Swapping the test-frame curvature field for a scalar microscope κ_{img} collapses the master-curve fit (pooled R^2 0.79 $\rightarrow$ -0.32) on the 62 microscope-covered specimens, despite the two channels agreeing as scalars ($r = 0.87$).

0.81). The sign of α is trivial (97% of specimens share it); the informative axis is magnitude, and one probe unlocks it. **One physical measurement of the target specimen—but nothing cheaper—recovers the scale.**

10 How many batches, and what to deliver

Batch power buys coverage, not point accuracy. The natural hope is that more training batches would eventually enable extrapolation. Using the validated surrogate (Fig. 2) to simulate nested prefixes of B batches, the answer is no for point prediction: the unseen-batch pooled R^2 plateaus near zero even at $B = 30$ (oracle-shape pooled -0.06, macro mean 0.12), because the irreducible between-batch log-scale spread $\tau = 1.04$ does not shrink with batch count, and adding the (weak) conditioning link does not lift the plateau. What *does* converge is the calibrated interval: the $\log|\alpha|$ predictive half-width falls from 2.47 at $B = 3$ to a floor of 2.23, reaching within 10% of that floor by $B^* \approx 5$ batches (Fig. 9a). This is the principled "how many batches" answer: enough to *calibrate* the between-batch spread ($\sim$5), after which additional batches buy coverage of the stiffness axis, not point accuracy. The simulation agrees with the real $B \leq 9$ learning curve, where a central target becomes predictable as batches are added but the stiffness-extreme batch 10 stays flat and negative regardless of count.

The deliverable: a calibrated interval for an unseen batch. If an unseen batch's stress cannot be predicted as a point, a usable model should return a *calibrated interval* and say so. We compare four methods for a target coverage of 0.90 (Table 8): pooled split-conformal, batch-level conformal, an empirical-Bayes hierarchical predictive band, and the surrogate posterior predictive.

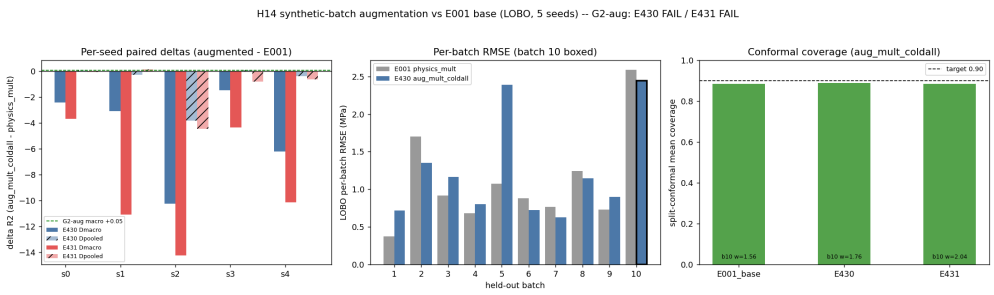


Figure 7: Synthetic-batch augmentation does not substitute for observing the scale. Paired per-seed change in unseen-batch macro-R^2 relative to the real-only baseline; augmentation is negative in all seeds and worsens as the synthetic spread is widened ($1.0\times \rightarrow 1.5\times$).

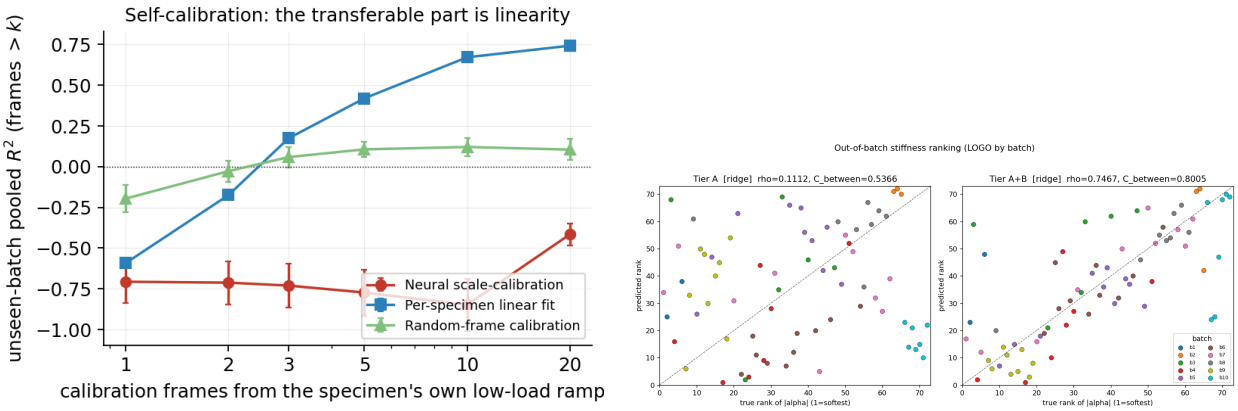
Table 7: Ranking specimens by stiffness out of batch (LOBO, ridge; permutation p over 1000 shuffles). Zero-measurement features rank at chance; one force–stroke probe ranks strongly. Tertile-accuracy chance is 0.33.

Feature set	Spearman ρ (95% CI)	perm. p	C-index (between-batch)	Tertile acc.
Tier A (zero-measurement)	0.11 $[-0.35, 0.57]$	0.16	0.54	0.35
Tier A + B (one probe)	**0.75** $[0.54, 0.86]$	**0.001**	**0.80**	**0.81**

Pooled conformal is tight in the median but collapses to 0.55 coverage on batch 10; batch-level conformal is valid but vacuously wide (median 8.9 MPa); the hierarchical band under-covers at low-signal points (batch 9, 0.14). Only the surrogate posterior-predictive band has coverage ≥ 0.80 on *all* 10 held-out batches (min 0.82 on batch 10) at a median width of 5.1 MPa—1.4× pooled conformal. The honest caveat is that batch 10 sits right at the threshold (0.82) and the band still misses its stiff tail (its Winkler score is the worst of the ten), so the interval is marginal exactly where it matters most; but it is the only method that does not fail outright on the hardest unseen batch. The correct output for a new batch is therefore this wide-but-honest interval, plus a request for one stiffness probe (§9) to sharpen it.

11 Discussion

The study yields a single coherent picture, now framed by what *is* recoverable. The curvature-to-stress map factorises into a batch-invariant shape and a per-specimen scalar stiffness; the shape is learnable and *transfers out of batch*, so from curvature alone one can predict what an unseen batch's stress field will look like up to its overall scale. The scale is a genuine material property spanning ∼30–50×—proven, using digitised per-specimen thickness, to be real modulus rather than a geometry artefact (t^3w geometry is only ∼7% of its variance)—and it is simply not present in the shape the specimen takes. Consequently no model capacity, no zero-measurement covariate (conditioning, geometry, thickness, microscope curvature, texture, shape descriptors), no domain-generalisation penalty, and no amount of generative augmentation adds anything on an unseen batch beyond a trivial linear SAM fit. The scale becomes available only through a measurement of the specimen itself—most cheaply, its own early load response, which a plain linear extrapolation converts into a good prediction ($R^2 = 0.74$), or a single force–stroke probe, which orders specimens by stiffness ($\rho = 0.75$). And because the between-batch spread is irreducible, batch count cannot buy point accuracy; what a designer can be handed for an unseen batch is a calibrated interval that



(a) Self-calibration frontier

(b) Stiffness ranking: Tier A vs. A+B

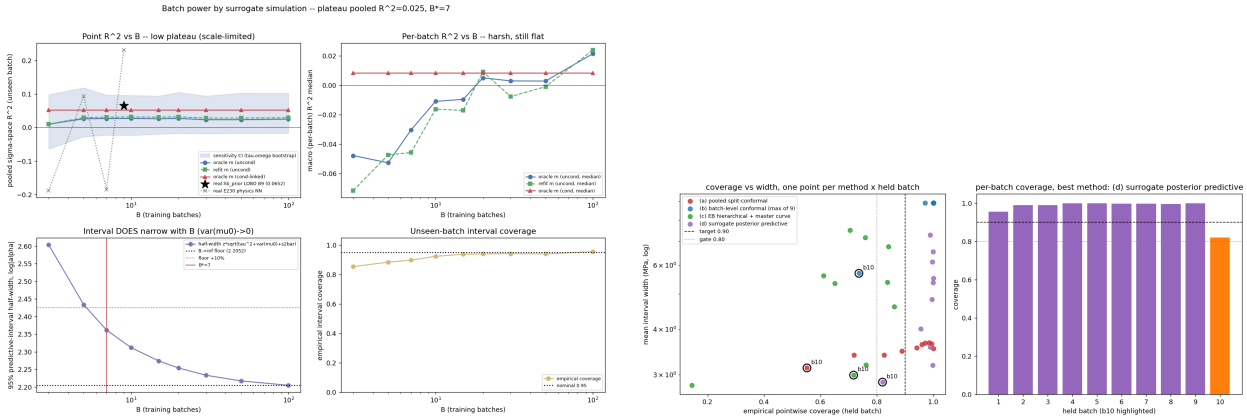
Figure 8: One measurement of the target specimen recovers the scale. (a) A per-specimen linear fit on the specimen's own first ∼10–20 low-load frames extrapolates well ($R^2 = 0.74$ at $k = 20$), while calibrating the neural scalar does not. (b) Predicted vs. true stiffness rank out of batch: zero-measurement Tier A is at chance ($\rho = 0.11$), one force–stroke probe ranks strongly ($\rho = 0.75$).

Table 8: Calibrated intervals for an unseen batch (target coverage 0.90; net predictions mean over 5 seeds). Only the surrogate posterior predictive clears 0.80 coverage on all 10 batches, at 1.4× the width of pooled conformal, which collapses on batch 10.

Method	Median cov.	Min cov. (batch)	Batch-10 cov.	Median width (MPa)	Batches < 0.80
(a) Pooled split-conformal	0.95	0.55 (b10)	0.55	3.55	2
(b) Batch-level conformal	1.00	0.74 (b10)	0.74	8.92	1
(c) EB hierarchical predictive	0.74	0.14 (b9)	0.72	5.37	7
(d) Surrogate posterior pred.	**1.00**	**0.82 (b10)**	**0.82**	5.11	**0**

covers all ten held-out batches, to be sharpened by one probe. This is a constructive, if sobering, design rule: read internal stress from shape *after* a one-off stiffness probe, and report an interval, not a point, until that probe is taken.

Two broader points. First, the result is a clean instance of the domain-generalisation lesson [10, 11]: under honest, grouped evaluation the elaborate methods do not beat the simple baseline, and the useful move is calibration/test-time adaptation [22, 23], not invariance. The materials framing is the genomics batch-effect one [19, 20]: correcting a between-group scale needs a measured covariate or replicated controls, which is exactly what conditioning time fails to provide and a stiffness probe supplies. Second, a methodological caution: an earlier iteration of this task reported a much rosier held-out score, produced with the locked test set used for early stopping and checkpoint selection and with specimen-level rather than batch-level splitting. Both inflate the estimate; under the leakage-free, batch-grouped protocol the within-batch score is ≈ 0.5 and the unseen-batch point score is near zero. We release the protocol and frozen splits precisely so that this class of error is not repeated.



(a) Coverage, not count

(b) Coverage vs. width by method

Figure 9: (a) Simulated batch power: unseen-batch point R^2 (left axis) stays near zero as batch count grows, while the calibrated interval half-width (right axis) reaches within 10% of its floor by $B^* \approx 5$. (b) Per-batch interval coverage vs. width for four methods; only the surrogate posterior-predictive band clears 0.80 coverage on every batch, including batch 10.

12 Limitations

The dataset is small in the way that matters—∼10 independent groups—so all estimates carry wide batch-level intervals and neural results wide seed intervals. Conditioning time is inferred per batch from a naming convention, is near-collinear with batch, and is unrecorded for batch 10, which additionally differs in camera and calibration (§6); it is treated as the hardest held-out batch, not the representative one. True microstructure (porosity, fibre volume fraction) was not measured and is the covariate most likely to index stiffness directly; the digitised thickness/width, microscope curvature and image texture we added are all genuinely zero-measurement but none predict the scale out of batch. The shape-transfer result carries a shuffled-shape-floor caveat (the field is dominated by a shared monotone trend); the calibrated interval is marginal on batch 10; and the augmentation result is a pre-registered null. Curvature is from imaged centrelines, not full-field; the stress field is taken as provided by the upstream pipeline (whose fixed-thickness section modulus we correct in the physics of §3 but do not re-derive); and temporal/rate effects are not modelled. The self-calibration and force–stroke results are Tier B and are reported as the practical upper bound, not as zero-measurement generalisation.

13 Conclusion

Across a wide, leakage-free sweep, the curvature-to-stress map for hygromorph flax/PLA factorises into a transferable shape and a per-specimen scalar stiffness: the shape transfers out of batch (scale-free per-specimen r^2 up to 0.91), while the scale—a genuine ∼30–50× material effect, not a geometry artefact—does not, and is recoverable from no zero-measurement covariate, no domain-generalisation method, and no generative augmentation, but from a single physical probe of the target specimen (self-calibration $R^2 = 0.74$; stiffness ranking $\rho = 0.75$). Because the between-batch scale is irreducible, additional batches buy calibration, not point accuracy; the honest deliverable for an unseen batch is a calibrated interval that covers all ten held-out batches at ≥ 0.80, sharpened by one probe. The path to a generally predictive model is broader stiffness coverage or a cheap per-specimen stiffness measurement, released here together with a reusable small-batch evaluation

benchmark.

Data and code availability

All processing, estimator-validation, diagnostic, model, evaluation, figure, and experiment-ledger code is in the project repository (`ann_v2/src/` and `ann_v2/src/experiments/`); the frozen splits, feature tables, per-experiment run directories, and the master table regenerate deterministically. The systematic study is reproduced by building the splits (`experiments.splits build`), running the experiment registry through `experiments.runner`, and rebuilding the ledger and master table (`experiments.ledger`, `experiments.build_master_table`).

Acknowledgements

The experimental hygromorph specimens and measurement conventions build on the flax-fibre biocomposite work of C. de Kergariou and colleagues at the Bristol Composites Institute.

References

[1] Charles M. Y. de Kergariou, Frédéric Demoly, Adam W. Perriman, Antoine Le Duigou, and Fabrizio Scarpa. The design of 4D-printed hygromorphs: State-of-the-art and future challenges. *Advanced Functional Materials*, 33(6):2210353, 2023. doi: 10.1002/adfm.202210353.

[2] Charles M. Y. de Kergariou, Antoine Le Duigou, Adam W. Perriman, and Fabrizio Scarpa. Design space and manufacturing of programmable 4D printed continuous flax fibre polylactic acid composite hygromorphs. *Materials & Design*, 225:111472, 2023. doi: 10.1016/j.matdes.2022.111472.

[3] Charles M. Y. de Kergariou, Hind Saidani-Scott, Adam W. Perriman, Fabrizio Scarpa, and Antoine Le Duigou. The influence of the humidity on the mechanical properties of 3D printed continuous flax fibre reinforced poly(lactic acid) composites. *Composites Part A: Applied Science and Manufacturing*, 155:106805, 2022. doi: 10.1016/j.compositesa.2022.106805.

[4] Charles M. Y. de Kergariou, Antoine Le Duigou, Victor Popineau, Victor Gager, Antoine Kervoelen, Adam W. Perriman, Hind Saidani-Scott, Giuliano Allegri, Túlio H. Panzera, and Fabrizio Scarpa. Measure of porosity in flax fibres reinforced polylactic acid biocomposites. *Composites Part A: Applied Science and Manufacturing*, 141:106183, 2021. doi: 10.1016/j.compositesa.2020.106183.

[5] Charles M. Y. de Kergariou, David Correa, Adam W. Perriman, and Fabrizio Scarpa. Effective material stiffness in curved actuators. *Advanced Intelligent Systems*, 8(2):2500668, 2026. doi: 10.1002/aisy.202500668. Introduces the Shape Actuation Modulus (SAM): SAM = $\Delta\sigma/\Delta(1/\kappa)$.

[6] Charles M. Y. de Kergariou, David Correa, Adam W. Perriman, Helmut Hauser, and Fabrizio Scarpa. A physical adaptive material motor unit neural network: a hygromorph composite material machine. *arXiv preprint arXiv:2606.18275*, 2026. doi: 10.48550/arXiv.2606.18275.

[7] Charles M. Y. de Kergariou, Hortense le Ferrand, Ali Momeni, Romain Fleury, Kunal Masania, Adam W. Perriman, and Fabrizio Scarpa. Towards engineering material neural networks. *arXiv preprint arXiv:2606.07262*, 2026. doi: 10.48550/arXiv.2606.07262.

[8] Pengcheng Xu, Xiaobo Ji, Minjie Li, and Wencong Lu. Small data machine learning in materials science. *npj Computational Materials*, 9:42, 2023. doi: 10.1038/s41524-023-01000-z.

[9] Ying Zhang and Chen Ling. A strategy to apply machine learning to small datasets in materials science. *npj Computational Materials*, 4:25, 2018. doi: 10.1038/s41524-018-0081-z.

[10] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations (ICLR)*, 2021. doi: 10.48550/arXiv.2007.01434.

[11] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *PMLR*, pages 5637–5664, 2021. doi: 10.48550/arXiv.2012.07421.

[12] Charles de Kergariou, Byung Chul Kim, Adam Perriman, Antoine Le Duigou, Sofiane Guessasma, and Fabrizio Scarpa. Design of 3D and 4D printed continuous fibre composites via an evolutionary algorithm and voxel-based finite elements: Application to natural fibre hygromorphs. *Additive Manufacturing*, 59:103144, 2022. doi: 10.1016/j.addma.2022.103144.

[13] Delphine Quereilhac, Thomas Caillault, Svetlana Terekhina, Mahadev Bar, Guillaume Morel, Marina Fazzini, Marwa Abida, Emmanuel De Luycker, and Pierre Ouagne. Development of a custom commingled flax/PLA wrapped yarn for additive manufacturing of long-fibre biocomposites. *Materials & Design*, 258:114628, 2025. doi: 10.1016/j.matdes.2025.114628.

[14] Melvin Josselin, Noëlie Di Cesare, Mickael Castro, Thibaut Colinart, Fabrizio Scarpa, and Antoine Le Duigou. Hygro-thermal coupling on 4D-printed biocomposites as key for meteosensitive shape-changing materials. *Virtual and Physical Prototyping*, 19(1):e2335233, 2024. doi: 10.1080/17452759.2024.2335233.

[15] Tarik Sadat. Machine learning-assisted tensile modulus prediction for flax fiber/shape memory epoxy hygromorph composites. *Applied Mechanics*, 4(2):752–762, 2023. doi: 10.3390/applmech4020038.

[16] Sivasubramanian Palanisamy, Nadir Ayrilmis, Kumar Sureshkumar, Carlo Santulli, Tabrej Khan, Harri Junaedi, and Tamer A. Sebaey. Machine learning approaches to natural fiber composites: A review of methodologies and applications. *BioResources*, 20(1):2321–2345, 2025. doi: 10.15376/biores.20.1.Palanisamy.

[17] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 7537–7547, 2020. doi: 10.48550/arXiv.2006.10739.

[18] Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Big data meets quantum chemistry approximations: The δ-machine learning approach. *Journal of Chemical Theory and Computation*, 11(5):2087–2096, 2015. doi: 10.1021/acs.jctc.5b00099.

[19] W. Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127, 2007. doi: 10.1093/biostatistics/kxj037.

[20] Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A. Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739, 2010. doi: 10.1038/nrg2825.

[21] Francesco Bandinelli, Lorenzo Peroni, and Alberto Morena. Elasto-plastic mechanical modeling of fused deposition 3D printing materials. *Polymers*, 15(1):234, 2023. doi: 10.3390/polym15010234.

[22] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *PMLR*, pages 9229–9248, 2020. doi: 10.48550/arXiv.1909.13231.

[23] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations (ICLR)*, 2021. doi: 10.48550/arXiv.2006.10726.

[24] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006. ISBN 026218253X.

[25] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 785–794, 2016. doi: 10.1145/2939672.2939785.

[26] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems 35 (NeurIPS), Datasets and Benchmarks Track*, 2022. doi: 10.48550/arXiv.2207.08815.

[27] Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, New York, 2007. ISBN 9780521686891. doi: 10.1017/CBO9780511790942.

[28] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01.

[29] Maziar Raissi, Paris Perdikaris, and George E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. doi: 10.1016/j.jcp.2018.10.045.

[30] George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021. doi: 10.1038/s42254-021-00314-5.

[31] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005. doi: 10.1007/b106715.

[32] Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101.

[33] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017. doi: 10.48550/arXiv.1712.00409.