CALIBRE: FOLD-CALIBRATED READOUTS FOR FROZEN PATHOLOGY ENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Frozen pathology encoders are evaluated on spatial gene expression prediction, but fixed readouts can obscure their value. HEST-Benchmark's ridge penalty leaves 256.0 of 256 effective degrees of freedom. We introduce CALIBRE, which adapts regularisation by selecting each encoder's penalty and projection dimension on training slides, averages seven encoders and applies an existing within-slide smoother. Calibration improves all seven pre-registered backbones on nine tasks. CALIBRE scores 0.4664 and the smoothed uncalibrated average 0.4575, versus 0.4480 from STFlow's published task scores. On 19 held-out breast cancer patients, calibration raises the primary encoder from 0.0811 to 0.1033 ($p = 0.0039$). CALIBRE scores 0.1213, versus 0.0765 for STFlow, a flow-matching model with training-selected epochs ($p = 0.00064$), or 0.0580 with a tuned learning rate ($p = 0.00042$). On a 22-patient kidney cohort from another institution and platform, registered before download, calibration improves the primary encoder for every patient and CALIBRE exceeds STFlow, 0.2144 versus 0.1795 ($p = 0.000009$). Predictors sharing encoder and smoother separate only against STFlow weakened by tuning. HER2ST's 0.2644 versus 0.2622 remains unresolved, and averaging is resolved on one of four cohorts. Training-selected frozen readouts change frozen-versus-trained comparisons.

1 INTRODUCTION

Predicting spatial gene expression from haematoxylin and eosin (H&E) histology tests what pathology foundation models learn about tissue. Predictors recover spot-level expression from morphology alone (He et al., 2020). Performance depends on both representation and readout. Task-trained predictors include end-to-end convolutional, multi-resolution and whole-slide models, some using frozen foundation-model features (He et al., 2020; Chung et al., 2024; Huang et al., 2025a; Zhu et al., 2025; Nonchev et al., 2025; Huang et al., 2025b; Lee et al., 2026). Alternatively, a small supervised readout maps frozen embeddings to expression. HEST-Benchmark and related evaluations compare encoders this way (Jaume et al., 2024; Chen et al., 2024; Padigela et al., 2025; Gindra et al., 2025; Ndubuisi, 2026), also supplying frozen baselines for trained predictors (Huang et al., 2025a). Readout specification affects both comparisons.

Readout capacity, regularisation and selection determine which predictive directions a representation reveals (Pimentel et al., 2020; Hewitt & Liang, 2019; Voita & Titov, 2020). Linear-probe choices can change conclusions about vision representations (Kolesnikov et al., 2019; Kornblith et al., 2019; Lee et al., 2023). Baseline tuning has reduced or reversed apparent progress in metric learning and recommendation (Musgrave et al., 2020; Ferrari Dacrema et al., 2019), while evaluation choices can rival leaderboard margins (Bouthillier et al., 2021; Dehghani et al., 2021). Frozen-versus-trained comparisons must distinguish representation limits from unsuitable readouts. Matched controls are essential because an overall gain neither locates its source nor establishes each component's necessity.

However, HEST-Benchmark fixes frozen encoders' projection dimensions and ridge penalties rather than selecting them on training slides. Its readout standardises embeddings, projects them to 256 principal components and sets the penalty from embedding width and gene count (Jaume et al., 2024). This gives $\alpha = 0.0078$, whereas the smallest retained squared singular value is about 2×10^2 across the 189 projected training fits we measured. Ridge retains 256.0 of 256 effective degrees of

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

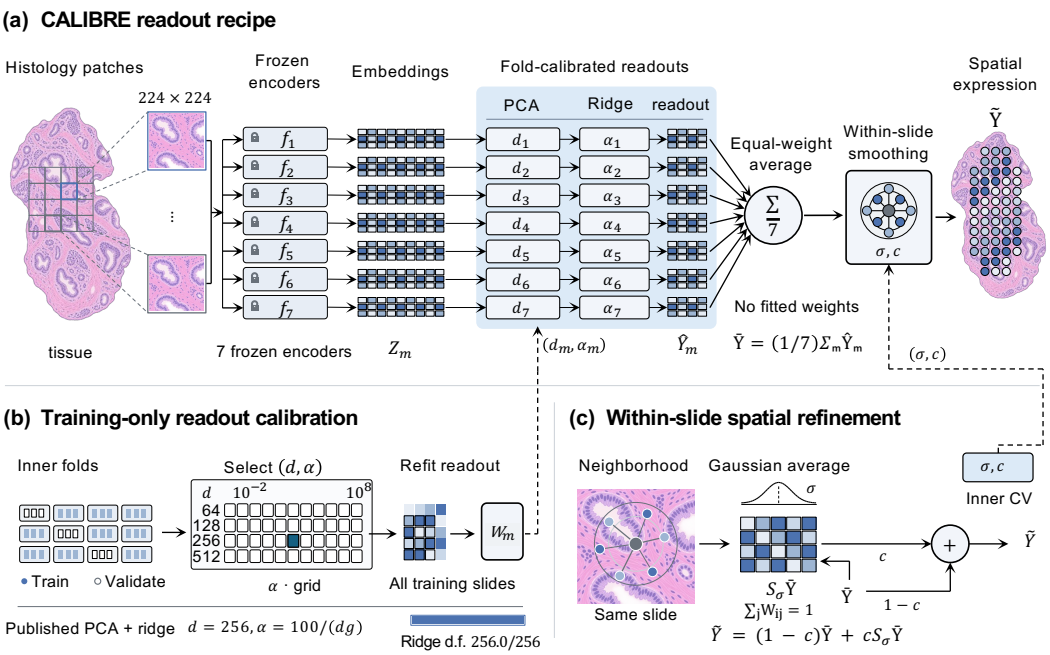


Figure 1: Readout calibration and the CALIBRE recipe. (a) Each of seven frozen encoders uses a separately selected principal component analysis (PCA)–ridge readout. Predictions are averaged with equal weight before within-slide smoothing. (b) Inner cross-validation on training slides selects each encoder's projection dimension d and ridge penalty α. Standardisation and PCA are refitted within each split, followed by a readout refit on all outer-training slides. The published setting, $d = 256$ and $\alpha = 100/(dg)$ for g genes, retains 256.0 of 256 ridge effective degrees of freedom. PCA still truncates rank. (c) An existing smoother mixes ensemble predictions with a row-normalised Gaussian average over spots within each slide. Bandwidth and mixing weight are selected on training slides after d and α are fixed. Held-out labels never enter readout or smoothing parameter selection. The H&E tissue and patch artwork is AI-generated and illustrative. Folds, matrices, spot patterns, the search cell and kernel profile are schematic.

freedom, applying negligible shrinkage. PCA still truncates rank and can regularise the fit (Dhillon et al., 2013). A common projection permits comparison under one protocol, but does not measure what each representation can achieve with a selected readout. Appendix L.1 traces the penalty's source and distinguishes this limitation from the rationale for a common projection. The trained model's paper scores its frozen baselines with a fixed random forest (Huang et al., 2025a). The readout's effect must be measured before attributing a performance gap to encoder adaptation.

To address this limitation, we propose CALIBRE, which calibrates each frozen encoder's PCA–ridge readout on training slides. It selects penalties and projection dimensions within training folds and refits each readout. Calibration adapts to embedding scale and spectrum without changing encoder weights. The full recipe averages seven predictions equally, fitting no combination weights, and adds existing post-estimation smoothing (P-ES) with training-selected parameters (Rolf et al., 2020). P-ES combines nearby predictions within slides (Figure 1). Matched controls give every arm the same smoother access. Encoders, folds and metrics remain fixed. Neither ridge nor the smoother is new.

Calibration raises all seven pre-registered backbones by $+0.0236$ to $+0.0397$ on the nine published HEST tasks. With P-ES available to both readouts, it raises H-Optimus-1 from 0.4377 to 0.4543 on HEST. On a held-out 19-patient cohort with fold-internal gene panels, it raises the same encoder from 0.0811 to 0.1033, and on a 22-patient kidney cohort registered before download it improves all 22 patients.

The full CALIBRE recipe reaches mean Pearson 0.4664, against 0.4480 for a whole-slide flow-matching model, averaged from its published task scores (Huang et al., 2025a), or 0.4313 against 0.415 over that paper's ten tasks. The released trained-model implementation selects epochs on

test folds. Against our rerun of that procedure, CALIBRE remains higher on nine of ten tasks. With epochs selected on training patients, CALIBRE leads by $+0.0448$ on the held-out cohort ($p = 0.00064$) and $+0.0349$ on the kidney cohort ($p = 0.000009$), and stays ahead when the trained model receives the same smoother. Learning-rate tuning lowered its ST-Net score, widening the margin to $+0.0633$. The trained-model comparison remains unresolved on HER2ST. After multiplicity correction, shared-encoder comparisons are also unresolved except against the weaker tuned model. Without calibration, the smoothed seven-encoder ensemble also exceeds the trained model, by $+0.0303$ on ST-Net ($p = 0.020$) and $+0.0337$ on the kidney cohort ($p = 6.7 \times 10^{-6}$), both post hoc and unadjusted, so averaging and smoothing carry most of the lead.

Our contributions are as follows:

- We show that the benchmark's ridge penalty applies negligible shrinkage, keeping df/d above 0.99998 in every fit (Figure 2c). Training-fold selection improves every encoder on HEST and on two held-out cohorts, including the pre-registered encoder with P-ES available to both readouts (Table O.1).
- We introduce CALIBRE, which calibrates each encoder, averages seven with equal weight and applies an existing training-selected smoother. It reaches mean Pearson 0.4664 on the nine published tasks, against 0.4480 averaged from the trained model's published task scores (Table 1).
- We compare frozen and trained predictors on held-out patients from two cohorts, one registered before download, making all choices, including gene panels, on training patients. CALIBRE exceeds the trained model on both, including with matched smoothing, while HER2ST and shared-encoder comparisons against the released model remain unresolved.

2 RELATED WORK

Frozen features and tuned baselines. HEST-Benchmark compares frozen pathology encoders through a shared readout (Jaume et al., 2024; Chen et al., 2024). Its appendix shows raw ridge reordering encoders, which it attributes to embedding width and avoids by projection (Appendix L.1). Selecting probe penalties per representation is routine in vision (Kornblith et al., 2019; Goyal et al., 2019). Probe capacity and controls affect representation comparisons (Hewitt & Liang, 2019; Voita & Titov, 2020; Pimentel et al., 2020), and linear-probe sensitivity is documented (Kolesnikov et al., 2019; Lee et al., 2023). Combining frozen pathology foundation models can improve weakly supervised classification (Neidlinger et al., 2026). Retuning baselines has also changed conclusions in metric learning, recommendation and language modelling (Musgrave et al., 2020; Ferrari Dacrema et al., 2019; Melis et al., 2018). We diagnose the inert penalty, test fold-internal selection and transfer, and measure frozen-versus-trained effects (Appendices X, X.1, X.2 and X.4).

Regularisation and fold-wise evaluation. Ridge error depends on representation spectrum and target alignment (Bordelon et al., 2020; Canatar et al., 2023). Projection also regularises (Frank & Friedman, 1993; Dhillon et al., 2013). CALIBRE excludes a target-centring component akin to batch correction (Appendices X.3 and W.3).

Trained spatial predictors. Task-trained predictors span convolutional, multi-resolution and whole-slide architectures (He et al., 2020; Chung et al., 2024; Huang et al., 2025a). They include neighbour-aggregation models (Pang et al., 2021; Zeng et al., 2022) and models that retrieve training expression through learned exemplars or contrastive embeddings (Yang et al., 2023; Xie et al., 2023). The comparator trains a whole-slide model on frozen UNI features. We rerun the comparator without claiming these architectures are unnecessary. Other methods change prediction losses, encoder training or post-hoc calibration of trained-model outputs (Wang et al., 2026), rather than the frozen readout. Appendix X.5 distinguishes these comparisons. Appendix X.7 records weight-release and protocol constraints on encoder-side comparisons.

Post-estimation smoothing. Smoothing predictions over a structural index is established (Rolf et al., 2020), including for spatial gene imputation (Sun et al., 2024). We measure its contribution with fold-selected parameters and destructive controls, rather than claim a new operation. Appendix X.6 distinguishes this setting from imputation and encoder-side calibration.

3 SETUP

Task and protocol. HEST-Benchmark contains ten tasks, nine on its leaderboard. The primary HEST analysis uses those nine, with the trained model's ten-task mean also reported. Each task, named by tumour type or organ (Figure 2a), pairs 224×224 histology patches with the fifty most variable genes at the same spots. The benchmark describes its folds as patient-stratified. Readouts fitted to frozen patch embeddings from training slides predict held-out expression. The seven pre-registered encoders are H-Optimus-1 and H-optimus-0 (Scalbert et al., 2026; Saillard et al., 2024), UNI2-h (Chen et al., 2024; MahmoodLab, 2025), Virchow2 (Zimmermann et al., 2024), Midnight-12k (Karasikov et al., 2025), Phikon-v2 and Phikon (Filiot et al., 2024; 2023). Performance is per-gene Pearson correlation within each fold, averaged over genes and folds. Data, folds, gene panels and metric are unchanged. Two inherited limitations affect every arm, including the trained comparator. The benchmark release carries no patient identifiers, so patient separation cannot be verified and our inner folds group by slide. It also fixes the gene panel once. HER2ST (Andersson et al., 2021), ST-Net and the kidney cohort carry HEST-1k patient metadata, so their splits hold out whole patients. Section-grouped inner folds lower CALIBRE's HER2ST score (Appendix G).

Readout fitting and parameter selection use training slides only. Smoothing uses held-out coordinates and predictions, never labels. A variant using test-derived feature normalisation is excluded from the headline results (Appendix W.2).

Published-row reproduction. Every compared backbone is checked against its published row on the benchmark's target. ResNet50, Phikon and H-optimus-0 reproduce scores and ranks with 0.0% relative deviation using their respective normalisation, pooling and constructor settings (Table E.1). A fourth backbone passes only after two extraction errors were corrected (Appendix W.1).

Trivial baselines. Pearson correlation is undefined for constant predictions. Shuffling embeddings within test folds gives -0.0037 over nine tasks. Predicting gene means from the cosine-nearest training slide in embedding space tests slide identity alone. It scores 0.0585, against the calibrated readout's 0.3888, on the seven tasks where it is defined (Appendix F).

Noise floor. Over three registered seeds, standard deviations are 0.00015 for CALIBRE's nine-task mean (0.00105 on its noisiest task), 0.00040 for the unsmoothed uncalibrated average, 0.00049 for the unsmoothed calibration contrast and 0.0014 for the trained-comparator rerun (Appendix I).

4 CALIBRE

CALIBRE combines readout calibration, equal-weight averaging and within-slide smoothing. Encoder weights remain frozen, and no combination weights are fitted.

Readout calibration. The published readout fixes the latent dimension at $d = 256$ and the ridge penalty at $\alpha = 100/(dg)$, which is 0.0078 at $g = 50$ genes. For a standardised training design projected onto d principal components with singular values $s_1 \geq \cdots \geq s_d$, ridge with summed squared loss has

$$\mathrm{df}(\alpha) = \sum_{j=1}^{d} \frac{s_j^2}{s_j^2 + \alpha} \tag{1}$$

effective degrees of freedom (Hastie et al., 2009). The s_j^2 grow with training-spot count and differ between encoders. A fixed α produces different shrinkage across settings. At the published α, df/d exceeds 0.99998 in all 189 encoder–task–fold fits, because even the 256th s_j^2 lies between 2.0×10^2 and 4.9×10^4. We select d and α on each fold's training slides using the benchmark metric and the grid in Appendix B, never a held-out slide. Selected cells usually retain more components than the published 256 ($d = 512$ in 157 of 189 fits) but fewer effective degrees of freedom, a median of 140 or 32% of d (interquartile range 20–50%), so calibration shifts regularisation from truncation to shrinkage. Selected penalties increase with training-spot count (Spearman $\rho = 0.73$), while df/d is only weakly related to training size (Appendix L.2). At the published (d, α) our implementation reproduces the published readout exactly, and its iterative solver has converged there (Appendix B).

The grid spans $d \in \{64, 128, 256, 512\}$ and eleven logarithmically spaced penalties from 10^{-2} to 10^{8}. Standardisation and principal component analysis are refitted within each training split, followed by ridge without an intercept. Inner validation holds out training slides. The selected configuration is refitted on all outer-training rows. With only one training slide, it uses two random halves of the spots and cannot assess cross-slide transfer. Encoder-specific calibration allows penalties and retained dimensions to differ across embedding widths.

Most of the gain does not require per-task selection. The cell most often selected by one backbone's inner grids, $d = 512$ and $\alpha = 10^4$, improves all fifteen tested backbones when applied unchanged across tasks. Selecting that cell with each scored task excluded still improves the seven pre-registered encoders (Appendix L and Figure L.1). A single calibrated encoder needs only its own encoding pass and CPU readout fits.

Prediction averaging. Each calibrated readout predicts every spot's gene expression. CALIBRE averages predictions equally across encoders without fitting combination parameters. Concatenation instead projects seven encoders jointly and weights each by its width. It is weaker under the published readout and does not separate from averaging after calibration (Appendix K).

Within-slide smoothing. CALIBRE uses existing post-estimation smoothing (Rolf et al., 2020). Within each slide, it mixes a spot's averaged prediction with a Gaussian-kernel average of that slide's predictions. Bandwidth is measured in median nearest-neighbour spacings. Once each encoder's d and α are fixed, bandwidth and mixing weight are chosen on the averaged predictions of the same inner folds, with ties favouring no smoothing. This step was registered after results for calibration and averaging were known. It adds $+0.0023$ to the nine-task mean at each of three seeds. Shrinkage towards the slide mean at the same weight costs 0.0063, and the same kernel over permuted coordinates costs 0.0050 (Appendix V).

For encoder m's calibrated predictions $\widehat{\mathbf{Y}}_{m,s}$ on slide s, the final output is

$$\bar{\mathbf{Y}}_s = \frac{1}{M}\sum_{m=1}^{M} \widehat{\mathbf{Y}}_{m,s}, \qquad \widetilde{\mathbf{Y}}_s = (1-c)\bar{\mathbf{Y}}_s + c\mathbf{S}_{\sigma,s}\bar{\mathbf{Y}}_s. \tag{2}$$

Here $M = 7$ and $\mathbf{S}_{\sigma,s}$ is the row-normalised Gaussian kernel over the spots of slide s, including self-weights and using identical weights for every gene. The search covers $c \in \{0, 0.1, \ldots, 1\}$ and $\sigma \in \{0.5, 1, 1.5, 2, 3, 5, 8\}$, and $c = 0$ recovers the calibrated ensemble exactly. Matched controls retain access to this operator. Coordinate permutation changes only its spatial information.

Excluded pooling component. A target-centring extension fits ridge to expression minus a selected fraction of each slide's mean. After calibration, its mean gain of $+0.0050$ over seven pre-registered backbones does not separate from zero ($p = 0.36$, task as unit). CALIBRE excludes this component (Appendix W.3).

5 RESULTS

Single-encoder calibration. Calibration raises the pre-registered encoder H-Optimus-1 from its published 0.4229 to 0.4465 on the nine-task mean, with gains on eight of nine tasks (one-sided Wilcoxon $p = 0.0059$). All seven pre-registered backbones improve, by $+0.0236$ to $+0.0397$ (Figure K.1 and Appendix Table A.2, with H-optimus-0 task values in Table B.2). With P-ES available to both readouts, calibration raises H-Optimus-1 from 0.4377 to 0.4543 on HEST, improving eight of nine tasks (two-sided exact Wilcoxon $p = 0.0195$). On HER2ST it raises the score from 0.2408 to 0.2697, improving five of eight patients ($p = 0.11$). Smoothing adds more to the published than the calibrated readout, $+0.0149$ against $+0.0077$ on HEST, indicating partial overlap (Appendix O). Calibrated rows cannot be ranked against uncalibrated leaderboard rows or baselines with different readouts (Lee et al., 2026). We make no such ranking (Appendix D).

Transfer of a fixed readout. The fixed cell of Section 4 improves all fifteen tested backbones and 132 of 135 backbone–task cells when applied unchanged. Averaged over backbones within tasks, it gains $+0.0368$, positive on all nine tasks. Shared slides prevent treating the backbones as fifteen

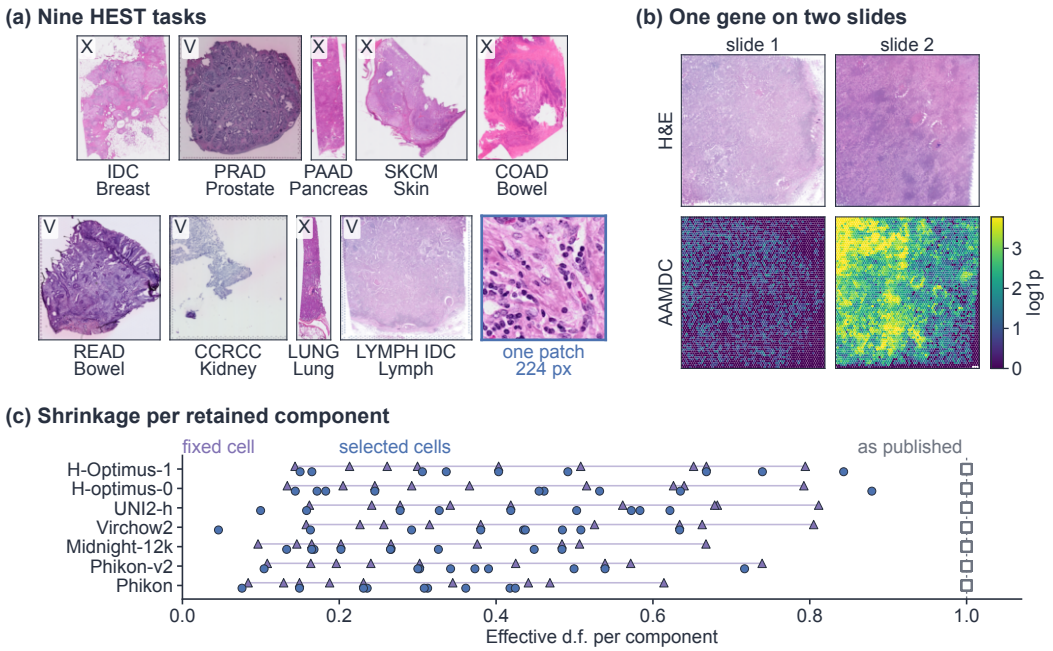


Figure 2: Nine tumour tasks and the readout penalty. (a) Tissue crops by task and organ, labelled X for Xenium and V for Visium, with one 224×224 patch at encoder resolution. (b) H&E and spot expression for two slides, on a shared scale clipped at the pooled second and ninety-eighth percentiles. The gene has the largest between-slide variance share (52.6%), and its slide means differ by 1.75 in $\log(1 + x)$ units. (c) Effective degrees of freedom per retained component, with one row per encoder and one mark per task. All sixty-three published-penalty cells coincide at the ceiling to the reported precision. At the fixed cell of Section 4, $d = 512$ and $\alpha = 10^4$, values span 0.08 to 0.81, roughly tenfold, and track training size. Circles show inner-selected cells averaged over outer folds, spanning 0.05 to 0.88.

independent replications. Task-excluded selection retains $+0.0240$ of the $+0.0287$ all-task gain over the seven encoders. Changing only the penalty at the published dimension recovers a mean 96% of the fixed-cell gain. It also improves both tested HESCAPE encoders (Gindra et al., 2025). A single-cell benchmark's probe selection changes one of four encoders (Appendices C and J).

A held-out patient cohort. An analysis of the breast cancer cohort of He et al. (2020) was registered before any encoding or scoring on it. Excluding four patients shared with HER2ST leaves 19 patients and 23,680 spots with patches. Each outer fold holds out all sections from one patient (Appendix Q). Calibration raises H-Optimus-1 from 0.0788 to 0.1024, improving 16 of 19 patients (two-sided exact Wilcoxon $p = 0.0024$, the registered primary test). All seven encoders improve. Averaging adds $+0.0088$, but this increment is not established ($p = 0.14$). Selecting the fifty genes within each training fold instead of cohort-wide raises H-Optimus-1 from 0.0811 to 0.1033, a gain of $+0.0222$, positive for 16 of 19 patients ($p = 0.0039$), and all seven encoders still improve. Subsequent registrations were written with knowledge of preceding results on these patients (Table Q.5).

Transfer across cohorts. The fixed HEST cell recovers only $+0.0045$ of the ST-Net cohort's $+0.0235$ per-fold gain (Table Q.4). The registered retained-fraction rule, $\mathrm{df}/d = 0.37$ from HEST, recovers $+0.0152$, improving 17 of 19 patients ($p = 0.0006$), but recovers only 30–67% of the per-fold gain across encoders (Appendices L.2 and Q). On the kidney cohort, where inner selection returns the HEST cell in 131 of 154 encoder folds, the two rules recover 84 to 121% of the per-fold gain (Appendix T), so whether a fixed rule suffices depends on the cohort.

Readout gains relative to encoder differences. MINT's transcriptomic pre-training (Lee et al., 2026) gains $+0.0251$ over H-optimus-0 (Appendix X.7), compared with $+0.0302$ from calibrating

Table 1: HEST mean Pearson over nine tasks and the trained model's ten. Frozen nine-task values average three seeds (one for UNI). Frozen ten-task values use seed one with P-ES. A dash denotes a readout not run on the tenth task. The published model and first rerun select epochs on test folds, as released. The second rerun follows the released schedule. Both reruns use three seeds (Appendix M). The last two rows score UNI under each paper's readout (Appendix R). Source papers report ten-task means to three decimals.

	nine published tasks		ten tasks
	no P-ES	P-ES	
H-Optimus-1, published readout	0.4228	0.4377	–
H-Optimus-1, calibrated	0.4465	0.4543	–
seven encoders, published readout	0.4524	0.4575	–
CALIBRE, seven calibrated encoders	0.4641	**0.4664**	**0.4313**
UNI, published readout	0.3857	0.3992	0.3677
UNI, calibrated	0.4171	0.4236	0.3918
whole-slide flow matching, published task scores (Huang et al., 2025a)	0.4480		0.415
rerun, epoch chosen on the test fold as released	0.4430		0.4102
rerun, fixed schedule, one test evaluation	0.3837		0.3534
strongest baseline in its table (Chung et al., 2024)	0.4328		0.395
UNI, the trained model's input, under its readout	0.3773		0.344
UNI under the leaderboard's readout	0.3873		–

that encoder's readout. Across fifteen encoders, calibration reverses eleven of 105 published pairs. None is resolved under both readouts when the nine tasks are resampled (Appendix H). On ST-Net it reverses 16 of 21 pairs, moving Midnight-12k from first to last, but no reversal survives Holm correction (Appendix Q).

Frozen-ensemble comparison. Without updating encoder weights, CALIBRE reaches 0.4664 against 0.4480 averaged from the whole-slide flow-matching model's published scores on nine tasks, and 0.4313 against 0.415 on its ten tasks. Against published task values, CALIBRE is higher on eight of nine tasks (two-sided exact Wilcoxon $p = 0.074$, task as unit). UNI, the model's input, scores 0.3773 in its table and 0.3873 on the leaderboard, a readout-only difference.

Comparator rerun and test-fold selection. The released STFlow procedure (Huang et al., 2025a) scores test samples after every epoch and keeps the best epoch per fold. Against a rerun on identical inputs with three seeds per arm under a registered rule, CALIBRE is higher on nine of ten tasks (one-sided Wilcoxon $p = 0.0049$, task as unit), losing only on CCRCC. Optimistically choosing each task's best fixed epoch on test folds gives 0.4197 and 0.3881, below CALIBRE on every task, and also choosing learning rates there leaves CALIBRE higher on eight of nine (Appendix M).

Matched component controls. Table O.1 crosses encoder count, calibration and smoothing access across three cohorts. Each arm selects its smoother on training data, with tasks or patients as the unit. With smoothing available, ensemble calibration adds $+0.0088$ on HEST, $+0.0134$ on HER2ST and $+0.0170$ on ST-Net. Averaging adds $+0.0121$ over the smoothed calibrated single encoder on HEST and $+0.0088$ on ST-Net, but subtracts 0.0052 on HER2ST. None of the four registered HEST/HER2ST contrasts survives Holm correction. With matched smoothing on ST-Net, only calibration survives (Appendices O and Q). Calibration has a positive mean in every matched contrast. Of four cohorts, only the kidney cohort resolves an averaging gain.

External comparison with the trained model. On HER2ST, which has coarser spots and an older platform than HEST, CALIBRE reaches 0.2644 with whole patients held out. STFlow's fixed schedule overfits. Test-patient epoch selection gives an optimistic 0.2742 (Appendix N). A denser training-side selector, frozen before its outer scoring, reaches 0.2622. CALIBRE leads by 0.0022 on five of eight patients, with a paired patient 95% interval of $[-0.0370, +0.0415]$, establishing neither superiority nor equivalence (Appendix P). On the ST-Net cohort, the same type of selector reaches 0.0765 with fold-internal panels under a registration preceding any of its training. CALIBRE leads by $+0.0448$, positive for 17 of 19 patients (two-sided exact Wilcoxon $p = 0.00064$, Holm-adjusted

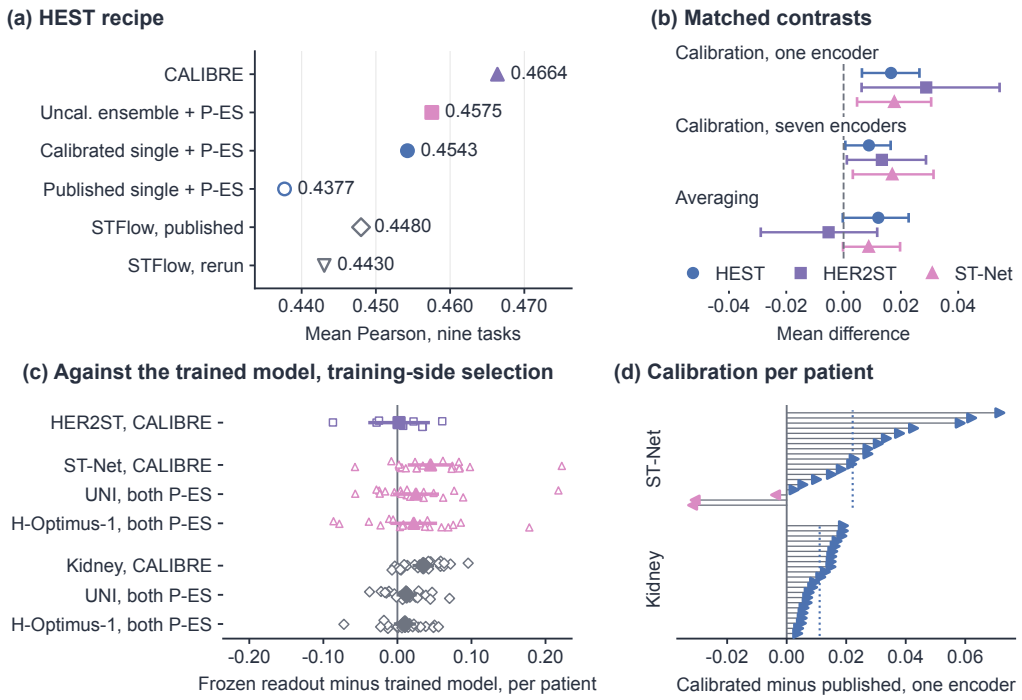


Figure 3: HEST performance, matched contrasts and the two held-out cohorts. (a) Nine-task HEST means, with frozen readouts averaged over three seeds with P-ES available and STFlow selecting epochs on test folds. (b) Matched contrasts with P-ES available to both arms on HEST tasks, eight HER2ST patients and nineteen ST-Net patients, with unadjusted 95% paired bootstrap intervals. (c) Frozen readouts minus STFlow with training-side selection at its released learning rate, which scored above the tuned one (Table 2). Open markers are patients and filled markers means with 95% t intervals. Rows marked both P-ES share encoder and smoothing. (d) Sorted unsmoothed calibrated-minus-published H-Optimus-1 differences per patient, with dotted means. ST-Net uses cohort-wide gene panels in (b), and both held-out cohorts use fold-internal panels in (c) and (d).

0.0019), meeting the registered rule. With STFlow given the same smoother, the margin is $+0.0388$ ($p = 0.0082$, Figure 3c and Table 2). Tuning STFlow's learning rate on inner folds lowers its score to 0.0580. CALIBRE leads by $+0.0633$ ($p = 0.00042$, Holm-adjusted 0.00084, Appendix S), with a positive mean difference whichever patient is omitted. A cohort-wide panel gave $+0.0416$ ($p = 0.0071$, Appendix R).

Frozen and trained predictors at matched input. STFlow uses UNI features (Chen et al., 2024). On those features, the calibrated and smoothed HEST readout reaches 0.4236, closing 65% of the gap between STFlow's frozen UNI row, 0.3773, and 0.4480 averaged from its published task scores. The readout is lower than the rerun only with test-fold epoch selection (Table 1). On the ST-Net cohort, with fold-internal panels and smoothing available to both predictors, the calibrated readout reaches 0.1075 on UNI and 0.1136 on H-Optimus-1, versus 0.0825 and 0.0915 for STFlow on the same features. Both differences are positive for 13 of 19 patients and unresolved with or without smoothing ($p = 0.0955$ and 0.13 smoothed, Holm-adjusted 0.19). The tuned model, 0.0225 lower with smoothing, trails the UNI readout by 0.0475 ($p = 0.011$). HER2ST's matched UNI comparison is unresolved ($p = 0.55$, Appendix R). The ST-Net advantage belongs to the full recipe, not a shared-encoder readout.

A second cohort registered before download. Every analysis of a 22-patient human kidney Visium cohort (Lake et al., 2023) was registered before any of its data were downloaded, with the ST-Net protocol unchanged (Appendix T). Calibration raises H-Optimus-1 from 0.1880 to 0.1991, higher for all 22 patients ($p = 4.8 \times 10^{-7}$), and improves all seven encoders. CALIBRE reaches

Table 2: ST-Net with gene panels chosen inside each training fold, means over 19 held-out patients. P-ES and STFlow's epochs, and for the tuned row its learning rate, are selected on inner patient folds, and STFlow averages three seeds. Each + P-ES column adds P-ES to the arm on its left. Contrasts give the mean difference, paired t interval, positive patients, two-sided exact Wilcoxon p and Holm correction within each registration (Appendix R lists the remaining secondary contrasts). The seven-encoder CALIBRE scores 0.1213.

encoder	published	calibrated	+ P-ES	STFlow	+ P-ES
UNI	0.0819	0.0982	0.1075	0.0765	0.0825
H-Optimus-1	0.0811	0.1033	0.1136	0.0853	0.0915
UNI, tuned STFlow				0.0580	0.0600

contrast	mean	95% interval	positive	p	Holm p
CALIBRE − STFlow, UNI	+0.0448	[+0.0163, +0.0733]	17/19	0.00064	0.0019
UNI, both with P-ES	+0.0250	[−0.0038, +0.0538]	13/19	0.0955	0.19
H-Optimus-1, both with P-ES	+0.0221	[−0.0071, +0.0513]	13/19	0.13	0.19
CALIBRE − tuned STFlow, UNI	+0.0633	[+0.0258, +0.1008]	17/19	0.00042	0.00084
UNI, both with P-ES, tuned STFlow	+0.0475	[+0.0067, +0.0883]	15/19	0.0108	0.0108
CALIBRE − STFlow + P-ES, H-Optimus-1	+0.0297	[+0.0004, +0.0590]	14/19	0.0361	
CALIBRE − STFlow + P-ES, UNI	+0.0388	[+0.0092, +0.0683]	15/19	0.0082	

0.2144 against 0.1795 for STFlow with training-selected epochs ($p = 0.000009$, Holm-adjusted 0.00003). With a shared encoder and smoother, the readout leads by +0.0124 on UNI ($p = 0.033$) and +0.0102 on H-Optimus-1 ($p = 0.050$), neither surviving Holm correction. Here averaging adds +0.0097 ($p = 0.00026$), whereas calibration inside the ensemble adds +0.0012 ($p = 0.32$).

Reduced budgets and registered extensions. The registered prediction that the frozen readout degrades less than the trained model as HER2ST training patients fall from seven to two fails. CALIBRE loses 0.0671 and the comparator 0.0450 (three of eight patients favour CALIBRE, one-sided $p = 0.875$) under the coarse training-side selector at one seed (Appendix U). Eight of nine registered extensions fail their rules. The ninth is CALIBRE's smoothing step (Appendices W and V).

6 LIMITATIONS

These results do not establish that trained architectures are unnecessary. The rerun comparator leads on one of ten tasks, the HER2ST comparison remains unresolved, and shared-encoder, matched-smoother comparisons are resolved only against a weaker tuned trained model. CALIBRE also costs more. Encoding 13,620 HER2ST spots through seven encoders took 1.8 GPU-hours, versus 0.08 for the trained model's single encoder (Appendix M). Averaging is resolved on one of four cohorts, calibration inside the ensemble is not established on the kidney cohort, and the registered annotation-budget prediction failed. The recipe was designed on HEST, so its HEST scores are not held-out results, and HEST lacks patient identifiers. Subsequent ST-Net registrations were written with knowledge of preceding results on the same patients, and a fixed penalty transfers to the kidney cohort but not to ST-Net.

7 CONCLUSION

The prescribed ridge penalty is inert. Training-fold readout selection improves every tested frozen pathology encoder by margins comparable to reported encoder gains, although no rank reversal survives multiplicity correction. With all choices made on training patients, CALIBRE exceeds a trained whole-slide model on two held-out cohorts, one registered before download, even with smoothing available to that model. Shared-encoder predictors separate only against a weaker tuned model. Matched controls do not establish general component superiority, and HER2ST leaves CALIBRE's advantage unresolved. Frozen encoders should therefore be compared through training-selected readouts.

REPRODUCIBILITY STATEMENT

The supplementary materials provide readout and smoothing routines, saved results, plotting inputs, and checks for matched-component and paired-selector analyses. Benchmark data, encoder weights and cached embeddings remain available from their original providers. Appendix B specifies the grids, inner folds, numerical conventions and selected configurations. Appendices O and P report independent-unit results and uncertainty calculations. Appendices Q and T describe construction of the two held-out cohorts and their analysis plans, the second registered before any of its data were downloaded. Appendix R reports trained-model comparisons at matched input. Appendix E specifies token pooling and embedding dimensions, with each checkpoint's repository in the accompanying code. Protocol deviations are disclosed in the relevant experiment sections. Appendix W.3 documents the partition defect affecting only the excluded pooling arm. Registered analysis plans were recorded under version control before execution and will be released with the repository. The supplied code supports the documented analyses, but not a complete replay of training from raw data.

ETHICS STATEMENT

This work uses only public, de-identified spatial transcriptomics data distributed with HEST-Benchmark and HEST-1k, including the HER2ST and ST-Net breast cancer cohorts and a human kidney cohort. Only the published comparator is trained, using its released implementation on those data with UNI and H-Optimus-1 features. The study evaluates frozen-encoder readouts and an evaluation protocol, not a clinical tool. Calibration changes score differences and some point-estimate encoder rankings, although no reversal survives multiplicity correction. These changes do not identify a clinically preferable encoder, which this benchmark was not designed to assess.

AI USE STATEMENT

Large language models supported code implementation, experimental design and analysis, manuscript drafting and copy-editing, and the illustrative H&E tissue and patch artwork in Figure 1. The authors directed the research, reviewed all AI-assisted work and verified every reported result against experimental outputs and cited sources. They take full responsibility for the manuscript.

REFERENCES

Alma Andersson, Ludvig Larsson, Linnea Stenbeck, Fredrik Salmén, Anna Ehinger, et al. Spatial deconvolution of HER2-positive breast cancer delineates tumor-associated cell type interactions. *Nature Communications*, 12:6012, 2021. doi: 10.1038/s41467-021-26271-2.

Pietro Balestra and Marc Nerlove. Pooling cross section and time series data in the estimation of a dynamic model: The demand for natural gas. *Econometrica*, 34(3):585–612, 1966. doi: 10.2307/1909771.

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119, pp. 1024–1034, 2020.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, et al. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems (MLSys)*, volume 3, 2021.

Leo Breiman. Stacked regressions. *Machine Learning*, 24(1):49–64, 1996. doi: 10.1007/BF00117832.

Abdulkadir Canatar, Jenelle Feather, Albert J. Wakhloo, and SueYeon Chung. A spectral theory of neural prediction and alignment. In *Advances in Neural Information Processing Systems*, volume 36, pp. 47052–47080, 2023. doi: 10.52202/075280-2038.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(70):2079–2107, 2010. URL https://jmlr.org/papers/v11/cawley10a.html.

Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024. doi: 10.1038/s41591-024-02857-3.

Youngmin Chung, Ji Hun Ha, Kyeong Chan Im, and Joo Sang Lee. Accurate spatial gene expression prediction by integrating multi-resolution features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11591–11600, 2024.

Colin Conwell, Jacob S. Prince, Kendrick N. Kay, George A. Alvarez, and Talia Konkle. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature Communications*, 15(1):9383, 2024. doi: 10.1038/s41467-024-53147-y.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8):1470–1480, 2024. doi: 10.1038/s41592-024-02201-0.

Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The benchmark lottery. *arXiv preprint arXiv:2107.07002*, 2021. URL https://arxiv.org/abs/2107.07002.

Paramveer S. Dhillon, Dean P. Foster, Sham M. Kakade, and Lyle H. Ungar. A risk comparison of ordinary least squares vs ridge regression. *Journal of Machine Learning Research*, 14(46): 1505–1511, 2013.

Tom Dupré la Tour, Michael Eickenberg, Anwar O. Nunez-Elizalde, and Jack L. Gallant. Feature-space selection with banded ridge regression. *NeuroImage*, 264:119728, 2022. doi: 10.1016/j.neuroimage.2022.119728.

Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. Are we really making much progress? a worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems (RecSys)*, pp. 101–109, 2019. doi: 10.1145/3298689.3347058.

Alexandre Filiot, Ridouane Ghermi, Antoine Olivier, Paul Jacob, Lucas Fidon, Axel Camara, Alice Mac Kain, Charlie Saillard, and Jean-Baptiste Schiratti. Scaling self-supervised learning for histopathology with masked image modeling. *medRxiv*, 2023. doi: 10.1101/2023.07.21.23292757.

Alexandre Filiot, Paul Jacob, Alice Mac Kain, and Charlie Saillard. Phikon-v2, a large and public feature extractor for biomarker prediction. *arXiv preprint arXiv:2409.09173*, 2024. URL https://arxiv.org/abs/2409.09173.

Ildiko E. Frank and Jerome H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993. doi: 10.1080/00401706.1993.10485033.

Ruben Gindra et al. A large-scale benchmark of cross-modal learning for histology and gene expression in spatial transcriptomics. *arXiv preprint arXiv:2508.01490*, 2025. URL https://arxiv.org/abs/2508.01490.

Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6390–6399, 2019.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2 edition, 2009. doi: 10.1007/978-0-387-84858-7.

Bryan He, Ludvig Bergenstråhle, Linnea Stenbeck, Abubakar Abid, Alma Andersson, Åke Borg, Jonas Maaskola, Joakim Lundeberg, and James Zou. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature Biomedical Engineering*, 4(8):827–834, 2020. doi: 10.1038/s41551-020-0578-x.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275.

Tinglin Huang, Tianyu Liu, Mehrtash Babadi, Wengong Jin, and Zhitao Ying. Scalable generation of spatial transcriptomics from histology images via whole-slide flow matching. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25550–25565. PMLR, 2025a.

Tinglin Huang, Tianyu Liu, Mehrtash Babadi, Rex Ying, and Wengong Jin. STPath: A generative foundation model for integrating spatial transcriptomics and whole-slide images. *npj Digital Medicine*, 8:659, 2025b. doi: 10.1038/s41746-025-02020-3.

Guillaume Jaume, Paul Doucet, Andrew H. Song, Ming Y. Lu, Cristina Almagro-Pérez, et al. HEST-1k: A dataset for spatial transcriptomics and histology image analysis. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks)*, volume 37, pp. 53798–53833, 2024. doi: 10.52202/079017-1704.

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.

kaiko.ai, Nanne Aben, Edwin D. de Jong, Ioannis Gatopoulos, Nicolas Känzig, et al. Towards large-scale training of pathology foundation models. *arXiv preprint arXiv:2404.15217*, 2024. URL https://arxiv.org/abs/2404.15217.

Mingu Kang, Heon Song, Seonwook Park, Donggeun Yoo, and Sérgio Pereira. Benchmarking self-supervised learning on diverse pathology datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3344–3354, 2023.

Mikhail Karasikov, Joost van Doorn, Nicolas Känzig, Melis Erdal Cesur, Hugo Mark Horlings, Robert Berke, Fei Tang, and Sebastian Otálora. Training state-of-the-art pathology foundation models with orders of magnitude less data. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 15967 of *Lecture Notes in Computer Science*, pp. 573–583, 2025.

Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1920–1929, 2019.

Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better ImageNet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2656–2666, 2019.

Blue B. Lake, Rajasree Menon, Seth Winfree, Qiwen Hu, Ricardo Melo Ferreira, Kian Kalhor, Daria Barwinska, Edgar A. Otto, Michael Ferkowicz, Dinh Diep, et al. An atlas of healthy and injured cell states and niches in the human kidney. *Nature*, 619(7970):585–594, 2023. doi: 10.1038/s41586-023-05769-3.

Jae-Hun Lee, Doyoung Yoon, ByeongMoon Ji, Kyungyul Kim, and Sangheum Hwang. Rethinking evaluation protocols of visual representations learned via self-supervised learning. *arXiv preprint arXiv:2304.03456*, 2023. URL https://arxiv.org/abs/2304.03456.

Minsoo Lee, Jonghyun Kim, Juseung Yun, Sunwoo Yu, and Jongseong Jang. MINT: Molecularly informed training with spatial transcriptomics supervision for pathology foundation models. *arXiv preprint arXiv:2603.07895*, 2026. URL https://arxiv.org/abs/2603.07895.

Jiabo Ma, Zhengrui Guo, Fengtao Zhou, Yihui Wang, Yingxue Xu, et al. A generalizable pathology foundation model using a unified knowledge distillation pretraining framework. *Nature Biomedical Engineering*, 10(3):545–564, 2026. doi: 10.1038/s41551-025-01488-4.

MahmoodLab. UNI2-h. https://huggingface.co/MahmoodLab/UNI2-h, 2025.

Gábor Melis, Chris Dyer, and Phil Blunsom. On the state of the art of evaluation in neural language models. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=ByJHuTgA-.

Yair Mundlak. On the pooling of time series and cross section data. *Econometrica*, 46(1):69–85, 1978. doi: 10.2307/1913646.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *European Conference on Computer Vision (ECCV)*, pp. 681–699, 2020. doi: 10.1007/978-3-030-58595-2_41.

Chimdi Walter Ndubuisi. What carries the signal in pathology foundation-model atlases? a patient-level controlled benchmark in breast cancer. *arXiv preprint arXiv:2608.00105*, 2026. URL https://arxiv.org/abs/2608.00105.

Peter Neidlinger, Omar S. M. El Nahhas, Hannah Sophie Muti, Tim Lenz, Michael Hoffmeister, et al. Benchmarking foundation models as feature extractors for weakly supervised computational pathology. *Nature Biomedical Engineering*, 10(6):1113–1123, 2026. doi: 10.1038/s41551-025-01516-3.

Kazuya Nishimura, Haruka Hirose, Ryoma Bise, Kaito Shiku, and Yasuhiro Kojima. Learning relative gene expression trends from pathology images in spatial transcriptomics. In *Advances in Neural Information Processing Systems*, volume 38, pp. 96393–96414, 2025. doi: 10.52202/085713-2900.

Kalin Nonchev, Sebastian Dawo, Karina Silina, Holger Moch, Sonali Andani, et al. DeepSpot: Leveraging spatial context for enhanced spatial transcriptomics prediction from H&E images. medRxiv preprint 2025.02.09.25321567, 2025. URL https://www.medrxiv.org/content/10.1101/2025.02.09.25321567v1.

Harshith Padigela, Shima Nofallah, Atchuth Naveen Chilaparasetti, Ryun Han, Andrew Walker, et al. PLUTO-4: Frontier pathology foundation models. *arXiv preprint arXiv:2511.02826*, 2025. URL https://arxiv.org/abs/2511.02826.

Minxing Pang, Kenong Su, and Mingyao Li. Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. *bioRxiv*, 2021. doi: 10.1101/2021.11.28.470212.

James D. Pearce, Sara E. Simmonds, Gita Mahmoudabadi, Lakshmi Krishnan, Giovanni Palla, et al. TranscriptFormer: A generative cell atlas across 1.5 billion years of evolution. *Science*, 393 (6806):aec8514, 2026. doi: 10.1126/science.aec8514.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. URL https://jmlr.org/papers/v12/pedregosa11a.html.

Tiago Pimentel, Naomi Saphra, Adina Williams, and Ryan Cotterell. Pareto probing: Trading off accuracy for complexity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3138–3153, 2020.

Esther Rolf, Michael I. Jordan, and Benjamin Recht. Post-estimation smoothing: A simple baseline for learning with side information. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1759–1769. PMLR, 2020.

Yanay Rosen, Yusuf Roohani, Ayush Agrawal, Leon Samotorčan, Tabula Sapiens Consortium, et al. Universal cell embedding provides a foundation model for cell biology. *Nature*, 656(8126):183–191, 2026. doi: 10.1038/s41586-026-10689-z.

Charlie Saillard, Rodolphe Jenatton, Felipe Llinares-López, Zelda Mariet, David Cahané, Eric Durand, and Jean-Philippe Vert. H-optimus-0. https://github.com/bioptimus/releases/tree/main/models/h-optimus/v0, 2024.

Marin Scalbert, Charlie Saillard, Thomas Peeters, Liam Gonzalez, Dasha Valter, Felipe Llinares-López, Zelda E. Mariet, and Rodolphe Jenatton. H-optimus-1: A foundation model for computational histopathology. *Cancer Research*, 86(8_Suppl):LB174, 2026. doi: 10.1158/1538-7445.AM2026-LB174.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, et al. Brain-Score: Which artificial neural network for object recognition is most brain-like? bioRxiv preprint 407007, 2018. URL https://www.biorxiv.org/content/10.1101/407007v1.

Eric D Sun, Rong Ma, and James Zou. SPRITE: improving spatial gene expression imputation with gene and cell networks. *Bioinformatics*, 40(Supplement_1):i521–i528, 2024. doi: 10.1093/bioinformatics/btae253.

Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023. doi: 10.1038/s41586-023-06139-9.

Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7:91, 2006. doi: 10.1186/1471-2105-7-91.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 183–196, 2020.

Changfan Wang, Xinran Wang, Donghai Liu, Fei Su, Lulu Sun, Zhicheng Zhao, and Zhu Meng. CHRep: Cross-modal histology representation and post-hoc calibration for spatial gene expression prediction. *arXiv preprint arXiv:2604.21573*, 2026. URL https://arxiv.org/abs/2604.21573.

Chuhan Wang, Adam S. Chan, Xiaohang Fu, Shila Ghazanfar, Jinman Kim, Ellis Patrick, and Jean Y. H. Yang. Benchmarking the translational potential of spatial gene expression prediction from histology. *Nature Communications*, 16:1544, 2025. doi: 10.1038/s41467-025-56618-y.

L. Weidener, M. Brkić, M. Jovanović, E. Ulgac, and A. Meduri. VCBench: A multi-dimensional benchmark for single-cell foundation models. bioRxiv preprint 2026.06.18.733146, 2026. URL https://doi.org/10.64898/2026.06.18.733146. Embeddings released at https://huggingface.co/datasets/appliedscientific/vcbench-embeddings.

Ross Wightman. PyTorch image models. https://github.com/huggingface/pytorch-image-models, 2019.

David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992. doi: 10.1016/S0893-6080(05)80023-1.

Ronald Xie, Kuan Pang, Sai W. Chung, Catia T. Perciani, Sonya A. MacParland, Bo Wang, and Gary D. Bader. Spatially resolved gene expression prediction from histology images via bi-modal contrastive learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 70626–70637, 2023. doi: 10.52202/075280-3095.

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, 2024. doi: 10.1038/s41586-024-07441-w.

Yan Yang, Md Zakir Hossain, Eric A. Stone, and Shafin Rahman. Exemplar guided deep neural network for spatial transcriptomics analysis of gene expression prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5028–5037, 2023. doi: 10.1109/WACV56688.2023.00501.

Yuansong Zeng, Zhuoyi Wei, Weijiang Yu, Rui Yin, Yuchen Yuan, Bingling Li, Zhonghui Tang, Yutong Lu, and Yuedong Yang. Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. *Briefings in Bioinformatics*, 23(5):bbac297, 2022. doi: 10.1093/bib/bbac297.

Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019. URL https://arxiv.org/abs/1910.04867.

Sichen Zhu, Yuchen Zhu, Molei Tao, and Peng Qiu. Diffusion generative modeling for spatially resolved gene expression inference from histology images. In *International Conference on Learning Representations (ICLR)*, pp. 19666–19690, 2025. URL https://openreview.net/forum?id=FtjLUHyZAO.

Eric Zimmermann, Eugene Vorontsov, Julian Viret, Adam Casson, Michal Zelechowski, et al. Virchow2: Scaling self-supervised mixed magnification models in pathology. *arXiv preprint arXiv:2408.00738*, 2024. URL https://arxiv.org/abs/2408.00738.

APPENDIX

The appendix separates method specification, component attribution and limits of generalisation. Appendices A and B define the comparison arms, fitting scope, selection, tie-breaking and computational cost. Appendices C to L assess whether calibration transfers beyond a selected grid and whether its gains exceed seed variability. They also cover statistical treatment, backbone configurations, trivial baselines, inherited exposures, the fifteen-encoder comparison, single-cell benchmark, encoder averaging and the fixed-cell analysis, including the penalty's origin. Appendices M to T distinguish full-recipe comparisons from matched-input and matched-component evidence across the trained comparator and three external cohorts. The remaining sections report annotation-budget sensitivity, CALIBRE's smoothing step, all excluded variants and negative results, and detailed related work.

CONTENTS

A Readout specification and arms

B Selection procedure

C The fixed penalty on a second histology benchmark

D Statistical treatment and primary comparisons

E Backbone configurations

F Trivial baselines not reported by the benchmark

G Two exposures inherited from the release

H Encoder comparisons across fifteen backbones

I The seed noise floor

J Readout calibration on a single-cell benchmark

K The size of the readout correction relative to adding encoders

L One constant, applied unchanged to fifteen backbones

M The trained comparator, rerun with and without test-fold selection

N Patient-held-out breast cohort

O Matched component controls with smoothing held available

P Denser training-side epoch selection for the external comparator

Q A held-out patient cohort

R The calibrated readout on the trained model's own encoder

S A training-only tuning budget for the trained model

T A second held-out cohort on a different platform

U Annotation-budget sensitivity on HER2ST

V Output smoothing on CALIBRE, with its parameters selected on inner folds

W Excluded variants and negative results

X Detailed comparisons with prior work

A READOUT SPECIFICATION AND ARMS

Pipeline. Seven frozen encoders map each patch to embeddings without further training. Each readout projects to d dimensions and fits ridge with penalty α. The published protocol fixes these at $d = 256$ and $\alpha = 100/(dg)$. CALIBRE selects both separately for each encoder by inner cross-validation on training slides only. The seven predictions are averaged equally and smoothed within each held-out slide. Bandwidth and mixing weight are also selected on training slides. Per-gene Pearson correlation is computed within the held-out fold, so centring occurs within that fold.

Readout arms. The same frozen embeddings support several readouts, whose results must be distinguished. Table A.1 lists their training-selected parameters and reporting locations. CALIBRE is the third row. The final two rows are excluded variants.

B SELECTION PROCEDURE

Fitting scope and tie-breaking. The scaler, projection and ridge are fitted only on the training rows of each split and refitted within every inner fold used for selection. No held-out row contributes to its transformation, which the released implementation checks. Every latent dimension is therefore compared under the same fitting scope.

Table A.1: Readout arms. The second machine in Appendix M was used for three trained-comparator tasks, not for any readout.

arm	selected on training slides	reported in
published readout	none, fixed at $d = 256$ and $\alpha = 100/(dg)$	Table A.2, Appendix L
calibrated readout	d and α	Tables 1 and A.2, Figure K.1, Appendices H, L and N
CALIBRE	d and α per encoder, an equal-weight average of seven, then a bandwidth and a mixing weight for the smoother	Table 1, Appendices M, N and V
ensemble before smoothing	d and α per encoder, then an equal-weight average of seven	Table 1, Figure K.1c, Appendices I, K, M, N and V
uncalibrated average	nothing, seven published readouts averaged	Table 1, Figure K.1c, Appendix K
calibrated single + P-ES	d and α for H-Optimus-1 or UNI, then its own smoothing parameters	Appendices O, Q and R
uncalibrated average + P-ES	seven published readouts averaged, then its own smoothing parameters	Appendix O
fixed cell	none, fixed at $d = 512$ and $\alpha = 10^4$	Appendices C and L
excluded pooling arm	d, α and a target-centring weight λ	Appendices W and W.3
transductive	excluded because it reads the test slide's statistics	Appendix W.2

Table A.2: Unsmoothed seed-one readouts for seven pre-registered backbones. Ranks are from the published leaderboard. Gains, wins and one-sided task-paired Wilcoxon tests compare calibration with the published leaderboard row. Calibration reverses one of twenty-one pairs, placing Phikon 0.0016 above Phikon-v2.

backbone	rank	published	calibrated	gain	wins	Wilcoxon
H-Optimus-1	1	0.4229	**0.4465**	+0.0236	8/9	0.0059
H-optimus-0	3	0.4150	**0.4452**	+0.0302	9/9	0.0020
UNI2-h	4	0.4141	**0.4391**	+0.0250	9/9	0.0020
Virchow2	6	0.4034	**0.4372**	+0.0338	9/9	0.0020
Midnight-12k	8	0.3952	**0.4296**	+0.0344	9/9	0.0020
Phikon-v2	15	0.3747	**0.4042**	+0.0295	9/9	0.0020
Phikon	19	0.3660	**0.4058**	+0.0397	9/9	0.0020

Exact ties select the smallest dimension, then the smallest penalty. The grid is evaluated in ascending order, and a cell replaces the current choice only when its score is strictly higher.

Search grids. Both parameters are selected jointly over $4 \times 11 = 44$ cells using training slides, never the held-out fold. The grids are fixed in advance and are not extended when a selection reaches an edge. At the upper penalty edge, predictions approach a known limit, which is scored directly as a floor (Appendix R).

- PCA dimension $d \in \{64, 128, 256, 512\}$. The published dimension is 256.
- Ridge penalty $\alpha \in \{10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3, 10^4, 10^5, 10^6, 10^7, 10^8\}$. The published penalty is $100/(dg)$, or 0.0078 at $d = 256$ and $g = 50$ genes.

Computational cost in ridge solves. An outer fold with K inner folds requires one ridge solve for the published readout and $44K + 1$ for calibration, or two hundred and twenty-one at $K = 5$. CALIBRE repeats calibration for each encoder. Averaging adds no solves. All readouts run on CPUs, with GPUs used only for patch encoding and comparator training. On the largest task, calibration

Table B.1: Task protocols from the benchmark splits. Ridge sums squared error over training spots, so a fixed penalty implies a different penalty per spot across tasks. The published 0.0078 corresponds to 6.0×10^{-6} to 1.2×10^{-7} per training spot, and the constant 10^4 to 7.7 to 0.15.

task	slides	outer folds	test slides per fold	training spots	test spots
CCRCC	24	6	4	55,849–67,651	6,569–18,371
PRAD	23	2	8, 15	25,891–36,819	25,891–36,819
LYMPH_IDC	4	4	1	14,972–14,976	4,988–4,992
LUNG	2	2	1	1,944–3,262	1,944–3,262
READ	4	2	2	3,819–4,588	3,819–4,588
SKCM	2	2	1	1,293–1,741	1,293–1,741
IDC	4	4	1	14,775–32,531	3,005–20,761
COAD	4	2	1, 3	5,079–10,572	5,079–10,572
PAAD	3	3	1	3,961–5,620	1,951–3,610

costs ten times the published readout's single fit. Solve counts are reported rather than wall-clock times because concurrent users maintained a shared-machine load of 47 to 56.

Inner folds. Inner validation holds out one training slide at a time. With only one training slide, as in LUNG, SKCM and one COAD fold, it uses two random halves of the spots. This fallback cannot assess cross-slide transfer. The exclusion below tests whether the gain persists without tasks whose folds all require that fallback. Excluding LUNG and SKCM, whose outer folds all have one training slide, calibration gains $+0.0227$ to $+0.0363$ across the seven pre-registered encoders, versus $+0.0236$ to $+0.0397$ over all nine tasks.

Numerical conventions. Embeddings are standardised column-wise on training rows, with unit scale for zero-variance columns. Unwhitened PCA retains $\min(d, n, p)$ components for n training rows and embedding width p. Thus, the three width-384 encoders retain their full width at $d = 512$. The PCA solver follows the library default for matrix shape, using randomised decomposition for wider encoders and the analysis seed as its random state. Ridge has no intercept. The final fit uses the LSQR solver of scikit-learn (Pedregosa et al., 2011) with its default tolerance of 10^{-4}, matching the benchmark. At the published cell this solver has converged. Replacing it with an exact Cholesky solve changes H-Optimus-1's nine-task mean from 0.42289 to 0.42288 and no task by more than 0.0002, so Equation 1 applies to the published fit. The inner grid uses exact Cholesky solutions. Inner scores average fold-wise mean per-gene Pearson correlations. P-ES uses the same criterion and folds. Bandwidth is measured in median nearest-neighbour spacings among the spots being smoothed. For a single-slide split, this uses the held-out half. Smoothing ties select the smaller mixing weight, then bandwidth. The supplementary archive fixes library versions and stores each fold's selected parameters alongside its results.

Seeds. Per-backbone analyses use three seeds on eight tasks and two on SKCM. Matched controls use three seeds on all nine HEST tasks and seed one on HER2ST (Appendix O). Encoders remain frozen and ridge is deterministic, so seeds affect only randomised singular value decomposition in PCA and inner-fold construction. The seed determining slide-identifier order is fixed. Variation before fixing that seed affects only the excluded pooling arm (Appendix W.3).

No published calibration reference. The reproduction gate compares the unmodified readout with each published row, agreeing within 0.1% for every pre-registered backbone. UNI, used only in matched comparisons, differs on IDC (Appendix R). This validates the implementation at the published dimension and penalty. No published counterpart independently constrains the calibrated readout. Both readouts nevertheless use the same implementation and folds, differing only in selected parameter values. An error specific to calibration would therefore have to affect another grid cell while leaving the published configuration correct.

Table B.2: H-optimus-0 task results and the d and α selected for each outer fold by the calibrated readout used in CALIBRE. Selections differ between folds on six of nine tasks, so a single parameter pair does not describe each task. Cross-validating the two fixed hyperparameters raises the nine-task mean by $+0.0302$.

task	published	tuned	d per fold	α per fold
CCRCC	0.2676	0.2909	512	$10^5, 10^6$
PRAD	0.3848	0.4266	128, 512	$10^5, 10^6$
LYMPH_IDC	0.2591	0.3012	128, 512	$10^5, 10^6$
LUNG	0.5590	0.6005	128	10^3
READ	0.2216	0.2611	512	10^4
SKCM	0.6454	0.6842	512	10^3
IDC	0.5976	0.6070	512	$10^4, 10^5$
COAD	0.3086	0.3344	256, 512	10^4
PAAD	0.4911	0.5008	64, 256, 512	$10^3, 10^4$
mean	0.4150	0.4452		

C THE FIXED PENALTY ON A SECOND HISTOLOGY BENCHMARK

A readout inherited from HEST-Benchmark. HESCAPE (Gindra et al., 2025) attributes its expression-prediction readout to HEST, describing evaluation *using a ridge regression head as performed in the* HEST *gene expression prediction benchmark*. It uses the same fifty most variable genes and patch size, with correlations averaged over genes and patient-level folds. The fifteen-page paper never specifies the ridge penalty, and the word ridge does not occur in its released code. The attribution suggests an inherited fixed penalty, but the released information does not establish its value.

The attribution also suggests that fixed readout settings can propagate between benchmarks. HESCAPE's headline finding that cross-modal pretraining *degrades* expression prediction is measured through that readout, and whether it holds under a calibrated readout is untested here.

The release has further omissions. We inspected one complete panel of 195 shards, 91 GB and 56,689 rows. No donor or patient identifier or sample accession is released anywhere in them. Three candidate identifier fields are constant across rows. The separately distributed splits use sample accessions and patient labels absent from those rows. Fitting directly to the release therefore pools patients across training and test, contrary to the intended splits. The supplied splits specify one partition of twelve, four and four samples. That partition cannot produce the paper's reported standard deviation across cross-validation folds, so the fold construction is also unavailable. Images are 1024×1024, rather than the reported 224×224, and the panel contains 343 genes from which fifty are selected.

A penalty comparison under a fully specified protocol. Reproducing the published correlations requires eight unspecified choices. Choosing them to match a reported score would not test reproduction. Our penalty comparison instead fixes these choices identically across arms. This prevents an arm-specific procedural advantage, but not interactions with the penalty. Targets are cell counts summed within patches and transformed by $\log(1+x)$. The fifty most variable genes are selected on training rows and reselected within every fold. Each tile is resized in full. Five folds use the sample index, the only released grouping, with sample separation explicitly checked. Two public encoders process all 56,689 patches, and each receives three readouts. The first uses the fixed HEST head that HESCAPE cites. The second cross-validates both hyperparameters on training samples only. The third uses the cell fixed on the first benchmark, unchanged and without reselection.

Calibration gains $+0.0174$ and $+0.0151$ (Table C.1), positive on all five folds for each encoder. The one-sided Wilcoxon value is 0.0312, its minimum at $n = 5$, and paired t-test p values are 0.0014 and 0.0008. Calibrated and fixed-cell rows coincide because all ten folds select the transferred cell, although this benchmark did not inform that cell. Across a different cohort, organ and assay, both encoder families select 10^4 at dimension 512, despite one encoder being 3.7 times larger. The ratio

Table C.1: Mean Pearson correlation on the second histology benchmark over five folds grouped by sample. Each encoder receives HEST's fixed head, a readout cross-validated on training samples, and the cell fixed on HEST without reselection.

encoder	readout	fold 1	fold 2	fold 3	fold 4	fold 5	mean
UNI, 1024-dim	the head as inherited	0.5273	0.5261	0.5064	0.4420	0.4987	0.5001
	cross-validated	0.5443	0.5526	0.5235	0.4583	0.5085	**0.5175**
	this paper's constant	0.5443	0.5526	0.5235	0.4583	0.5085	**0.5175**
GigaPath, 1536-dim	the head as inherited	0.5402	0.5568	0.5205	0.4730	0.4861	0.5153
	cross-validated	0.5560	0.5738	0.5368	0.4804	0.5050	**0.5304**
	this paper's constant	0.5560	0.5738	0.5368	0.4804	0.5050	**0.5304**

to the prescribed penalty is 1.28×10^6, independently recovering the same six-order difference as on HEST. The second encoder shows that this result is not specific to the first.

The inner criterion distinguishes grid cells. Across forty-four cells, the best-minus-worst score ranges from 0.103 to 0.125 on every fold of both encoders. The selected cell exceeds the fixed published cell by 0.017. Best inner scores of 0.388 to 0.453 are below outer scores of 0.458 to 0.574. Inner folds score within-sample variation in one held-out training sample, whereas outer folds pool four samples and also score between-sample variation. Selection therefore measures different variation from outer scoring, yet recovers the transferred cell. As on HEST (Figure L.1), the inner surface is nearly flat across the six grid penalties below the optimum, peaks, then collapses two decades above it.

Half-precision encoding varies with batch boundaries. Re-encoding the same panel with the same encoder at a different batch size changes all $56{,}689$ embeddings by up to 5% of their median magnitude. The mean gain changes by 3×10^{-6}, no fold-level arm changes by more than 9×10^{-6}, and no fold changes its selected cell. Thus, a 5% embedding perturbation changes only one fold's fourth decimal, without altering the comparison's conclusion.

This comparison neither reproduces HESCAPE's published rows nor corrects its leaderboard. HESCAPE's missing probe, folds and patient identifiers prevent either. Three public encoders have reported correlations with between-fold standard deviations there, whereas HEST reports no such deviation for any row. We evaluate two encoders on one of six panels because the release contains 1.24 TB of pixels without cached features, requiring image encoding for every compared model. The unreleased sample-to-accession mapping is recoverable only as a hypothesis and is not used. Released sample indices suffice to keep these folds sample-disjoint. Under a protocol stated in full, the comparison establishes that the fixed HEST head, which HESCAPE cites, applies too little shrinkage on the second benchmark's data. It also establishes that the constant transfers to a benchmark that had no part in choosing it.

The evidence covers two encoders on one panel of one additional benchmark, under our reconstruction of its protocol. It shows that the correction transfers across a cohort, organ and assay. It does not show which penalty HESCAPE itself used, or that the correction transfers universally.

D Statistical treatment and primary comparisons

Primary comparisons. The primary HEST comparison is CALIBRE against the trained comparator rerun as released, with tasks as the test unit (Appendix M). Secondary comparisons comprise the fixed-schedule rerun and four HER2ST verdicts, each assessed by its registered rule (Appendix N). Each held-out-cohort registration specifies its own primary test. These cover calibration with five Holm-adjusted secondaries, retained-fraction transfer, fold-internal gene panels and three trained-model comparisons, with Holm correction also reported across the latter (Appendices Q, R, S and T). The matched-component diagnostic uses four prespecified two-sided exact Wilcoxon tests with Holm correction and unadjusted paired bootstrap mean intervals (Appendix O). The dense-epoch diagnostic reports unadjusted paired patient intervals and a descriptive rule (Appendix P). Other tests are exploratory and uncorrected for multiplicity. Failure to reach a threshold leaves the effect unresolved at the test's resolution rather than establishing its absence. Differences are computed from unrounded means and can differ in the last digit from differences of printed means.

Table E.1: Pooling and embedding dimensions for fifteen backbones, including the seven in Table A.2. CLS denotes the class token. Checkpoint repositories and preprocessing settings accompany the code. Virchow2 concatenates CLS with the mean of patch tokens starting at index five, matching both its model card and the benchmark. Midnight-12k uses CLS only, matching the benchmark but not its model card. The arm following the model card was therefore voided. Phikon-v2 also uses CLS. The additional backbones are GPFM (Ma et al., 2026), the Kaiko models (kaiko.ai et al., 2024), Lunit ViT-S/8 (Kang et al., 2023) and ResNet50 (He et al., 2016).

backbone	pooling	width
H-Optimus-1	CLS	1536
H-optimus-0	CLS	1536
UNI2-h	CLS	1536
Virchow2	CLS $\oplus$ mean	2560
Midnight-12k	CLS only	1536
Phikon-v2	CLS	1024
Phikon	CLS	768
GPFM	CLS	1024
Kaiko ViT-B/8	CLS	768
Kaiko ViT-B/16	CLS	768
Kaiko ViT-L/14	CLS	1024
Kaiko ViT-S/8	CLS	384
Kaiko ViT-S/16	CLS	384
Lunit ViT-S/8	CLS	384
ResNet50	average pool	1024

Patient-level tests and intervals take each held-out patient's score as the unit and condition on the fitted selection pipeline. Leave-one-patient-out fits share most training patients, so these intervals do not include variability from rerunning selection on independent training sets.

No leaderboard ranking. Calibrated rows are not ranked against published leaderboard rows. Such a ranking would compare selected readouts with unselected ones, and the published rows have no uncertainty intervals.

Exploratory multiplicity and a resolution floor. Reporting all tests discloses the analytical alternatives but does not correct for their multiplicity. Familywise or false-discovery adjustments could be applied jointly to the exploratory tests and the per-backbone column of Table A.2. The abstract's held-out-cohort p values lie outside those families, and each is a registered primary. On ST-Net they are the fold-internal-panel calibration test, $p = 0.0039$, and the first trained-model comparison of each of two registrations, 0.0019 and 0.00084 after Holm correction within each. On the kidney cohort the trained-model comparison is 0.00003 after Holm correction over three (Appendices Q, R, S and T). The nine-of-nine Wilcoxon result, $p = 0.0020$, is exactly $1/512$, the minimum at $n = 9$. It adds no information beyond all nine task differences being positive. Failed tests are retained, including the registered annotation-budget prediction (Appendix U).

E BACKBONE CONFIGURATIONS

The void arm in Appendix W.1 used model-card pooling rather than the benchmark's extraction rule. This doubled the embedding width and raised the score 4.31% above the published row. Appendix W.1 reports a second configuration error found by the same check. Table E.1 documents the backbone-specific settings. Fourteen of fifteen backbones use their resolved encoder-specific normalisation without an additional resize. Lunit ViT-S/8 is the exception. Omitting its resize and crop changes READ performance by -2.34%, so it uses the transform of timm (Wightman, 2019). The 0.9 crop ratio resizes to 248 before centre cropping, rather than leaving the image unchanged. All reported widths were measured from stored embeddings, not model cards.

Table F.1: Trivial baselines. The constant predictor uses training gene means. The nearest-slide predictor uses gene means from the cosine-nearest training slide. Shuffling permutes embeddings within the test fold. Constant predictions give a zero Pearson denominator. Their scores are therefore undefined, rather than failed, and appear as n/a.

task	constant	nearest slide	shuffled	tuned
CCRCC	n/a	0.0488	0.0013	0.2909
COAD	n/a	−0.0388	−0.0002	0.3344
IDC	n/a	0.0236	−0.0001	0.6070
LUNG	n/a	n/a	−0.0152	0.6005
LYMPH_IDC	n/a	0.0811	0.0007	0.3012
PAAD	n/a	0.0683	0.0061	0.5008
PRAD	n/a	**0.2345**	−0.0017	0.4266
READ	n/a	−0.0079	−0.0058	0.2611
SKCM	n/a	n/a	−0.0181	0.6842
mean	n/a	0.0585	**−0.0037**	0.4452

F TRIVIAL BASELINES NOT REPORTED BY THE BENCHMARK

Table F.1 provides three simpler baselines beside the tuned H-optimus-0 readout, all under the benchmark's folds and metric. The nearest-slide and shuffled baselines use H-optimus-0 embeddings.

The shuffled score of -0.0037 shows that performance requires the original embedding–target pairing, rather than the metric, fold structure or target marginal distribution alone. It does not bound leakage. Permutation destroys both morphology and any embedding covariate that proxies the label, so a near-zero result cannot distinguish them. Slide, site, platform, stain and coordinate leakage require specific controls that were not run (Appendix G). The nearest-slide predictor recovers 15% of the calibrated score, 0.0585 versus 0.3888, over the seven tasks where it is defined. Comparing it with the nine-task mean of 0.4452 would instead give 13%. The two omitted tasks are among the three highest-scoring, so matching tasks produces the larger share. Dataset identity does not explain most of the score, but this does not establish that morphology explains the remainder. Images also encode site, stain and acquisition, which the unrun controls in Appendix G would need to separate. PRAD has the highest nearest-slide score, 0.2345. Its folds contain eight and fifteen slides, making slide identity informative and visible to the metric. LUNG and SKCM each contain two slides, leaving one training slide per fold. Their nearest-slide predictions are constant and therefore undefined.

G TWO EXPOSURES INHERITED FROM THE RELEASE

Patient exposure not bounded by these measurements. The benchmark provides no patient identifiers, so patient separation cannot be verified and is not claimed. Comparing selected parameters with a fixed cell cannot bound patient leakage. A fixed penalty still fits predictions to training data. Ridge can learn patient-specific information whenever training and test slides share a patient, and penalties can exploit that information differently.

Across the fifteen backbones in Appendix L, cross-validation gains a mean $+0.0003$ over the fixed cell, spanning $[-0.0061, +0.0072]$ and negative for six backbones. Both arms share the exposure, so this difference does not bound it. The closest available evidence is the second benchmark (Appendix C), whose checked sample-disjoint folds all select the same constant across ten folds. Patient separation is unverified there too.

HER2ST provides patient identities and directly tests part of this exposure. Its sensitivity arm groups inner folds by section rather than patient, allowing selection to use sister sections from the held-out section's tumour. This lowers the unsmoothed calibrated ensemble from 0.2515 to 0.2458 on held-out patients and lowers six of seven encoder scores (Appendix N). Thus, inner-fold exposure lowers the score on this cohort. It says nothing about outer-fold exposure from patients shared between training and test slides.

These results do not establish patient-disjoint folds. Controls for slide, site, platform, stain and coordinate leakage were not run.

The released files cannot establish whether inner-fold selection crosses patient boundaries or whether selection of the fifty target genes used the whole dataset.

No patient identifier is distributed. Patient identifiers are absent from all three possible sources. Split tables contain only sample identifiers, patch paths and expression paths. Per-sample annotations contain spot coordinates, tissue flags and count summaries. No metadata table is released. The description of patient-stratified folds therefore cannot be independently checked by downstream users, whose finest available grouping is the slide. Our inner folds use that grouping. Multiple training slides from one patient could fall on opposite sides of an inner split, inflating a hyperparameter's apparent quality through correlated tissue.

Gene panels reselected within training splits. The fifty most variable genes define each task. Reselecting them within training splits changes the target and prevents comparison with the twenty-six published rows, so benchmark scores retain the official panel. A registered repeat on the held-out ST-Net cohort instead selects fifty genes within each training fold. Calibration gains $+0.0222$, versus $+0.0235$ with the cohort-wide panel (Appendix Q).

Slides within a task need not share a gene space. COAD's four 541-gene panels overlap on 412 genes, while CCRCC includes slides with $17{,}943$ and $36{,}601$ genes. We select and retrieve the fifty targets by name. Every target is present in every slide across all nine tasks. Concatenating full matrices by column position could silently align different genes.

Implications for between-arm comparisons. Published and calibrated readouts use the same folds and genes, so both inherit these exposures. This removes a simple additive shift in their difference, but does not exclude unequal exploitation. The published readout selects no parameters, whereas calibration selects two. Shared folds and genes rule out a readout-specific additive advantage, but not every effect of these exposures on the calibration gain. The fixed cell in Appendix L requires no deployment selection and recovers a mean 98% of the per-task selection gain.

H Encoder comparisons across fifteen backbones

Eligibility fixed before measurement. Eligibility required a construction specified by the benchmark's encoder library rather than a model card, weights available without a licence gate, and a published rank outside those covered by the initial seven. Eight encoders qualified, at ranks 13, 16, 18, 20, 21, 22, 23 and 26, extending coverage from the top nineteen published ranks to the top twenty-six. Two others were excluded for dependencies, not results. One required a separate fork of the model library, and the other's bundled-module licence was unsuitable for an anonymised archive. Both exclusions preceded any calibrated score from the eight encoders. All eight reproduce their published rows. The largest per-task deviation is 0.152%, below the 0.25% gate calibrated on the initial seven.

The initial seven encoders were pre-registered before embedding extraction, whereas the additional eight are post hoc.

Reordering under the calibrated readout. The top five published rows span 0.0168 across the twenty-six-encoder leaderboard. The calibrated readout averaged by CALIBRE reverses eleven of the hundred and five pairs among the fifteen tested encoders. Per-task calibrated scores appear in Appendix L. Resampling nine tasks yields calibrated-gap intervals excluding zero for two reversed pairs. No reversed pair has both its published and calibrated gap resolved by that resampling.

I The seed noise floor

The 0.0019 registered margin in Appendices N, P and W is the largest across-seed deviation of a three-task mean pooling increment over calibration. It was measured across seven pre-registered backbones using five seeds (Appendix W.3). Here we measure the ensemble's own seed variability.

Table I.1: Unsmoothed nine-task means of the calibrated ensemble and the uncalibrated average under three seeds, with their spread across seeds and the largest per-task standard deviation.

	calibrated ensemble, unsmoothed	uncalibrated average	difference
seed 1	0.4641	0.4520	+0.0121
seed 2	0.4639	0.4528	+0.0111
seed 3	0.4642	0.4525	+0.0117
standard deviation	0.00015	0.00040	0.00049
range	0.00030	0.00078	0.00098
largest task standard deviation	0.00075, SKCM	0.00253, READ	0.00245, READ

Seed variability of the ensemble. The unsmoothed calibrated ensemble and uncalibrated average were evaluated under seeds two and three on all nine tasks, totalling 252 readouts (Table I.1). The design and interpretations of both possible outcomes were registered before execution. Recomputation matched both stored seed-one scores within 5×10^{-7} before results were reported. Seeds affect only randomised projection and the random spot halves used when slide splits are unavailable.

With smoothing, CALIBRE scores 0.4664, 0.4662 and 0.4665 across three seeds, with standard deviation 0.00015. The largest per-task standard deviation is 0.00105, on SKCM. Nine-task variability is roughly an order of magnitude below the decision margin, about one twelfth of 0.0019. Treating this variability and the trained rerun's 0.0014 standard deviation as separate processes gives a difference standard deviation of 0.0014, about fifteen times smaller than the +0.0211 margin. The unsmoothed calibrated ensemble exceeds the rerun on eight of nine tasks under every seed, with $p = 0.0098$ each time and mean margins of +0.0209 to +0.0212. Against the uncalibrated average, the unsmoothed calibrated ensemble is higher on eight of nine tasks under seeds one and three, at $p = 0.0098$. Seed two improves seven tasks, at $p = 0.0137$, with CCRCC becoming negative at −0.0023. The decisions retain the registered 0.0019 threshold. Three extensions fall between observed variability and that threshold, namely shared weights on CALIBRE at +0.0016, the rank constraint at +0.0012 and the penalty profile at +0.0007. None separates at task level either ($p = 0.18$, 0.37 and 0.37). Seed deviations measure execution variability on fixed tasks, not uncertainty across independent tasks or patients.

J READOUT CALIBRATION ON A SINGLE-CELL BENCHMARK

A single-cell benchmark tests whether readout sensitivity extends beyond HEST. Its cells are split randomly without grouping. The experiment tests only whether selecting its previously unvalidated probe hyperparameters changes the comparison.

The benchmark and its probe. VCBench (Weidener et al., 2026) compares four frozen single-cell foundation models, TranscriptFormer, UCE, scGPT and Geneformer (Pearce et al., 2026; Rosen et al., 2026; Cui et al., 2024; Theodoris et al., 2023), by predicting a protein-derived target from RNA embeddings. Its probe fits ridge directly to raw embeddings with penalty 1.0, without projection, standardisation or parameter selection. Three source locations specify this configuration. Embedding widths are 512, 1152, 1280 and 2048, so the same penalty spans a fourfold difference in feature count. For the widest embedding, scikit-learn reports an ill-conditioned design at $\mathrm{rcond} = 1.86 \times 10^{-8}$.

The incumbent is reproduced to 7.8×10^{-9} before any comparison. All four published scores are recovered from the released embedding banks and public source data. This reproduction establishes three properties. Embeddings use the random $80/20$ fallback split, not the site-held-out split. The target is the fifty principal components of the antibody matrix stored in the released annotations. This explains the reported protein count of 50, compared with the source's 134 antibodies. The projection is not refitted and is identical across arms, preventing an arm-specific procedural advantage but not an interaction. Its fitting rows are undocumented, so training-only fitting is not claimed. Reproduction also confirms embedding–target row alignment. Precision checks varying one factor at a time attribute the entire score discrepancy to embedding precision, with no effect from target

Table J.1: Probe scores on VCBench for four single-cell foundation models. The published probe fits ridge with penalty 1.0 to raw embeddings. The tuned probe selects projection dimension and penalty by five-fold cross-validation on training cells, and the last column gives its choice.

encoder	width	published	tuned	gain	selected
TranscriptFormer	2048	0.2318	0.2318	$+0.0000$	the incumbent's cell
UCE	1280	0.1315	0.1361	$+0.0045$	raw, $\alpha = 0.1$
scGPT	512	0.0638	0.1306	$\mathbf{+0.0668}$	$d = 512$, $\alpha = 0.01$
Geneformer	1152	0.0007	0.0008	$+0.0001$	$d = 512$, $\alpha = 0.01$

precision. The released 32-bit embeddings reproduce scores within 7.8×10^{-9}, whereas 64-bit embeddings differ by 4.8×10^{-6}. All comparisons therefore retain the released 32-bit representation.

Cross-validating the probe. Five-fold cross-validation on training cells selects projection dimension from $\{\text{none}, 64, 128, 256, 512\}$ and penalty over eleven decades. The grid includes the published cell. Reproduction at that cell is checked for every encoder before selection. Table J.1 reports both probes.

Transfer of the calibration effect. Three encoders improve and the leader does not. The largest gain, scGPT's $+0.0668$, is comparable with adjacent published gaps of 0.0631, 0.0677 and 0.1002. It exceeds the smallest gap but falls 0.0009 short of the second, which scGPT would need to cross. Calibration gives scGPT 0.1306 versus UCE's published 0.1315, leaving the ranking unchanged. Its selected dimension is 512 from width 512, so the projection is full-rank. The gain comes from standardisation and penalty selection, not dimension reduction. The published probe does not standardise scGPT's features, which have unequal scales.

Limits of the four-encoder comparison. The leader's fixed configuration is already optimal on the tested grid, giving exactly zero gain. A setting not selected by validation need not be suboptimal. Geneformer reaches only 0.0008, because its embedding carries almost no signal for this target and no readout in the grid recovers it. A foundation-model score of 0.0007 alone is therefore not evidence of a defective probe.

K The size of the readout correction relative to adding encoders

Comparisons in Figure K.1. The uncalibrated average exceeds the published trained comparator, and calibration increases the margin. Panel (a) compares each of seven pre-registered backbones with its own published nine-task mean. All improve. Appendix H extends the analysis to fifteen encoders. Panel (b) compares the rank-one encoder's calibrated score with the best of twenty-six published rows by task, including its one loss. It measures the scale of the correction, not encoder order, because only our readout is cross-validated. Panel (c) plots unsmoothed means against encoder count under both readouts, with the trained whole-slide model shown as a dashed line. One calibrated encoder remains below that line, two exceed it, and the published readout requires four.

This analysis expresses the readout correction in units of the gain from adding encoders, relating it to encoder-side improvements. The comparison and its falsifier were fixed before any of these ensemble results were measured.

Design. All seven frozen encoders have cached embeddings, so combining them requires no additional encoding. Their separately calibrated per-spot predictions are averaged equally. Every combination analysis here is before smoothing. The same averaging is applied to published readouts. Both readouts produced results for all nine tasks and seven encoders without failures.

Correcting the readout and adding encoders overlap. Relative to the rank-one encoder's published nine-task mean of 0.4229, calibrating one encoder gains $+0.0236$ under CALIBRE's readout. Averaging seven published readouts gains $+0.0291$, and combining both changes gains $+0.0412$. This is 78% of the $+0.0527$ expected from independent effects, indicating overlap of about one fifth.

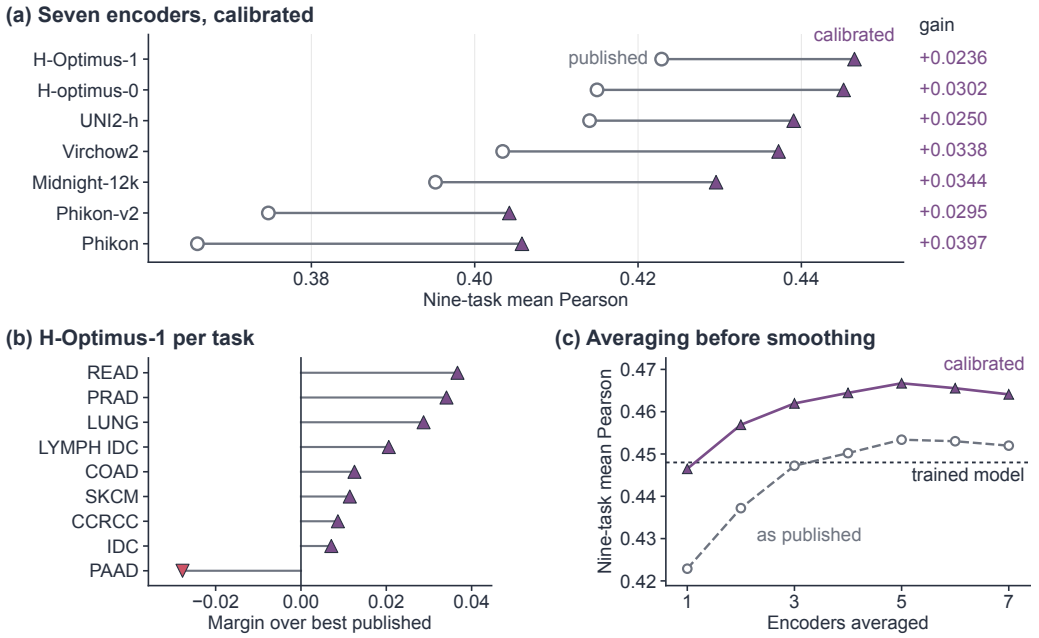


Figure K.1: Unsmoothed HEST seed-one results. (a) Published nine-task means (grey circles), calibrated means (purple triangles) and gains for the seven pre-registered backbones. (b) Calibrated H-Optimus-1 versus each task's best of twenty-six published encoders. It is higher on eight tasks and lower on PAAD. This comparison does not rank encoders. (c) Nine-task means as encoders are averaged under published and calibrated readouts, both unsmoothed, versus 0.4480 averaged from the trained model's published task scores. The unsmoothed single encoder remains below that score, while its smoothed counterpart exceeds it (Appendix O).

Table K.1: Nine-task means as encoders are added in published-rank order to the rank-one encoder, under the published readout and under the calibrated readout before smoothing.

encoders added	0	1	2	3	4	5	6
published readout	0.4229	0.4372	0.4472	0.4502	**0.4534**	0.4530	0.4520
calibrated, before smoothing	0.4465	0.4569	0.4620	0.4644	**0.4667**	0.4656	0.4641

Performance across ensemble sizes. The seven-encoder average is one point on an ensemble-size curve. Encoders are added in fixed published-rank order, not selected for their observed benefit. Table K.1 starts from the rank-one encoder's published readout and gives the calibrated curve beside it.

Both curves peak after four additions and then decline, as equally weighted averaging includes the fifteenth- and nineteenth-ranked encoders. The maximum averaging gain is $+0.0305$, versus $+0.0291$ for the pre-registered seven-encoder configuration beyond the peak. All reported method results retain those seven encoders. The maximum after four additions is exploratory, not a planned configuration.

Bounds on learned combination weights. Learned weights, as in stacking (Wolpert, 1992; Breiman, 1996), could improve on equal averaging. Fitting them on other outer folds would use held-out labels and violate the evaluation protocol. Inner-training-fold fitting would avoid this leakage, but was not run. Stored predictions allow a bound for one class of stackers by fitting a shared weight vector to each fold's own held-out target. Least squares uses the fold-centred target but does not exactly optimise the reported metric. It maximises correlation for a single centred target, confirmed numerically within 3×10^{-16}. Here, one shared weight vector determines a mean of fifty separately normalised gene correlations. In a constructed four-gene example, direct maximisation

Table K.2: Nine-task means of the seven published readouts under equal weights and under two oracle weightings fitted by least squares on each held-out fold.

combination of the seven	published readout
equal weights, as run	0.4520
one weight vector per fold, fitted on the held-out fold	0.4626
one per gene per fold, fitted the same way	0.5102

exceeds least squares by $+0.0157$. The least-squares solution in Table K.2 is therefore a lower bound on that oracle, not its optimum.

Equal weighting is not close to optimal. An oracle with seven free parameters per fold adds $+0.0106$ under the published readout, some of which inner-fold stacking might recover. The $+0.0236$ single-encoder correction under CALIBRE's readout is 81% of the $+0.0291$ from equally averaging six additional encoders. Per-gene weights reach 0.5102, gaining $+0.0873$ over the rank-one row, relative to which calibration contributes 27%. Per-gene stacking is a valid, less constrained family not covered by the shared-weight bound. It is reported as a comparator, not the reference for calibration. Its seven parameters per gene across fifty genes in each fold are fitted on the same rows used for scoring.

Feature concatenation and prediction averaging. Feature concatenation joins all seven embeddings into $10{,}496$ dimensions before the same readout. Concatenation was registered before prediction averaging. Averaging was registered after concatenation trailed on one task, although the rationale recorded at that point did not depend on the observed ordering. A shared decomposition retains maximum-variance directions, while encoders of the same patches largely agree. Differences between encoders may have low variance and be discarded, although this mechanism was not measured. Such differences are what combination seeks to exploit. After column standardisation, concatenation also weights encoders by width. Virchow2 contributes 24% of columns and Phikon 7%.

Under the published readout, concatenation scores 0.4336 versus 0.4520 for prediction averaging, which is higher on all nine tasks (two-sided sign test, $p = 0.0039$). Calibration narrows this difference. The fixed $d = 256$ readout is weakest where feature space is largest, and selecting d recovers most of concatenation's deficit. Under the calibrated readout averaged by CALIBRE, before smoothing, concatenation scores 0.4606 versus 0.4641. It is lower on six of nine tasks (one-sided Wilcoxon $p = 0.125$), so the methods do not separate. Concatenation retains at most 512 components in total, while averaging retains up to 512 per encoder. Neither comparison fixes total rank. Costs also differ. Concatenation requires 22.6 CPU-hours of $10{,}496$-dimensional decompositions, including 14.7 for CCRCC alone. Averaging reuses cached readout results.

Scope and cost of encoder averaging. The seven-encoder average is not entered on a leaderboard intended for single frozen encoders. The unsmoothed calibrated ensemble was fixed before the final four of its seventy readouts completed. Averaging seven published readouts already reaches 0.4520 without calibration. The computational costs differ. Combining encoders requires seven image-encoding passes, whereas CALIBRE's calibration uses $44K + 1$ CPU ridge solves per encoder for K inner folds on cached features, without a GPU (Appendix B). This distinction concerns cost, not capability.

L ONE CONSTANT, APPLIED UNCHANGED TO FIFTEEN BACKBONES

A fixed cell selected on one backbone. The published penalty is 0.0078. Across twenty-seven outer folds, the most frequent inner-selected cell uses penalty 10^4 and dimension 512. Applying it unchanged to every task gains $+0.0197$ on the nine-task mean, or 83% of the rank-one encoder's $+0.0236$ per-task selection gain. It exceeds per-task selection on four of nine tasks. The cell changes both penalty and dimension, since the published dimension is 256. The next analysis separates their contributions. The constant shows that the published penalty is about six orders of magnitude too small. Placing it within the optimum's decade recovers most of the per-task gain.

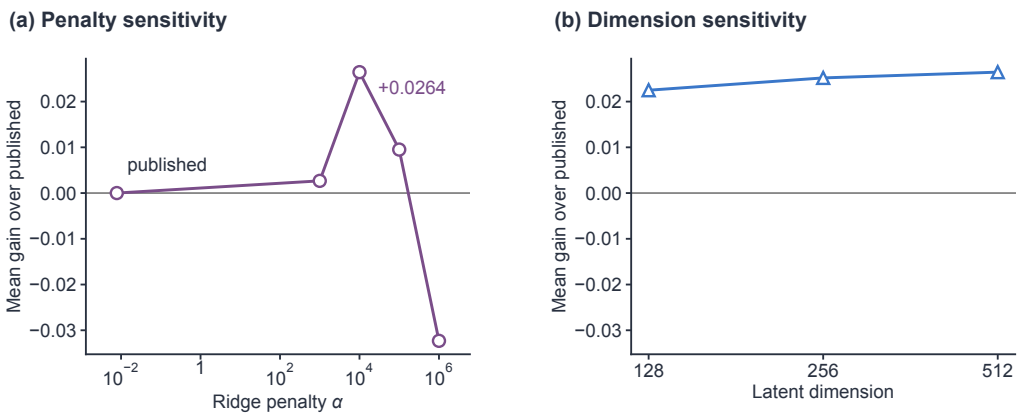


Figure L.1: Penalty and dimension sensitivity for H-Optimus-1 on the six tasks included in the computationally feasible sweep. (a) Mean gain over the published readout versus penalty at the swept dimension. The published penalty has zero gain by construction. The optimum is about one decade wide. A decade below the best cell retains $+0.0027$ of $+0.0264$, and a decade above retains $+0.0095$. Two decades above performs worse than no change. (b) Dimension sensitivity at the best penalty, on the same vertical scale.

Penalty sensitivity and the narrow optimum. On the six tasks included in the computationally feasible sweep, dimension has little effect. At the same penalty, dimensions 128, 256 and 512 recover 103%, 115% and 121% of the per-task gain. Penalty sensitivity is much greater (Figure L.1). One decade below the constant retains 12% of that gain, one above retains 44%, and two above gives -0.0323, worse than no change. This narrow optimum explains the remaining value of per-task selection. Selection has its largest advantage over the constant on PRAD and CCRCC, with twenty-three and twenty-four slides, and a smaller advantage on LYMPH_IDC. PRAD and CCRCC provide the most inner groups for locating the peak. On two-slide tasks, the constant gains half again as much. Across all seven pre-registered encoders and nine tasks, changing only the penalty at the published dimension of 256 retains a mean 96% of the fixed-cell gain, ranging from 86% to 105% by encoder.

Fourteen additional backbones. The $+0.0197$ gain and 83% recovery above concern one backbone. The same fixed cell is evaluated on fourteen others using cached embeddings, without parameter selection or gradients.

Origin and scope of the fixed cell. Penalty 10^4 and dimension 512 were selected once as the modal cell of the rank-one backbone's twenty-seven inner grids. All backbones below use that pair without reselection. The other fourteen are out of sample with respect to encoder identity, but not data, so applying the cell to them tests representation transfer within one benchmark. All fifteen share nine tasks. Taking the mode across twenty-seven outer grids allows a fold's held-out slides to influence it through other folds' training sets. On a second cohort, organ and assay, all ten folds independently select the unchanged constant (Appendix C), testing data transfer. That result does not remove selection exposure from the HEST estimate. The task-excluded analysis below addresses that exposure.

All columns of Table L.1 are nine-task means. The untouched and per-task reference arms are recomputed, not copied, enabling the reproduction check below. They differ from Table A.2's published leaderboard column by at most 0.0003. The share column divides the fixed-cell gain by the same backbone's per-task gain. The beats column counts tasks where the fixed cell exceeds per-task selection.

Results across fifteen backbones. The constant improves all fifteen backbones by a mean $+0.0368$, with range $[+0.0197, +0.0545]$, and improves 132 of 135 backbone–task cells. The backbone sign test gives 3.1×10^{-5}, or 2^{-15}, but is not used for inference. Fifteen encoders sharing

Table L.1: Fifteen backbones under the untouched readout, per-task calibration and the fixed cell ($d = 512$, $\alpha = 10^4$) selected once on the rank-one backbone. Gain is the fixed cell's improvement over the untouched readout.

backbone	rank	untouched	per-task	one cell	gain	share	beats
H-Optimus-1	1	0.4229	0.4465	0.4426	+0.0197	83%	4/9
H-optimus-0	3	0.4150	0.4452	0.4411	+0.0261	86%	4/9
UNI2-h	4	0.4138	0.4391	0.4372	+0.0234	93%	4/9
Virchow2	6	0.4033	0.4372	0.4300	+0.0268	79%	4/9
Midnight-12k	8	0.3952	0.4296	0.4277	+0.0325	95%	2/9
GPFM	13	0.3793	0.4159	0.4199	+0.0406	111%	6/9
Phikon-v2	15	0.3747	0.4042	0.4083	+0.0336	114%	5/9
Kaiko ViT-B/8	16	0.3735	0.4172	0.4179	+0.0444	102%	4/9
Lunit ViT-S/8	18	0.3678	0.4043	0.4078	+0.0400	110%	5/9
Phikon	19	0.3660	0.4058	0.4049	+0.0389	98%	5/9
Kaiko ViT-B/16	20	0.3645	0.4143	0.4110	+0.0464	93%	4/9
Kaiko ViT-L/14	21	0.3642	0.4117	0.4111	+0.0469	99%	5/9
Kaiko ViT-S/8	22	0.3513	0.4009	0.4056	+0.0544	110%	6/9
Kaiko ViT-S/16	23	0.3493	0.3977	0.4038	+0.0545	113%	6/9
ResNet50	26	0.3253	0.3537	0.3498	+0.0245	86%	4/9

nine tasks, slides and twenty-seven folds are not independent observations. Averaging encoders within tasks gives $+0.0368$, positive on all nine tasks (one-sided Wilcoxon $p = 0.0020$). This is the minimum at $n = 9$ and reflects the common sign only. Mean recovery of the per-task calibration gain is 98%, and median recovery is 98%. Six backbones recover more than 100%, outperforming per-task selection with a fixed pair. The constant wins 68 of 135 backbone–task comparisons despite selecting no parameters separately for those tasks.

Task-excluded constant selection. For each scored task, the fixed cell is the mode of the rank-one backbone's inner grids over the other eight tasks. It remains $(512, 10^4)$ for five tasks and becomes $(512, 10^5)$ for READ, IDC, COAD and PAAD. The overall mode is weak, occurring in eight of twenty-seven folds. Across seven pre-registered encoders, task-excluded selection gains $+0.0240$ versus $+0.0287$ for all-task selection, positive for all seven. On the selecting rank-one encoder, the gains are $+0.0124$ and $+0.0197$. Excluding the scored task retains about five sixths of the gain. The remaining gain reflects benchmark-level selection of the reported constant.

Limits of the fixed-cell interpretation.

- The constant requires no deployment selection but originates from one backbone's inner grids. At a fixed penalty, three dimensions recover 103%, 115% and 121% of the per-task gain, indicating that the transferred decade matters more than the exact cell. This narrows, but does not eliminate, the selection. Discrete choices can overfit at any granularity.
- The arm changes two parameters. The published dimension is 256 and the fixed cell uses 512. The dimension sweep attributes most of the gain to the penalty.
- Encoder-score spread changes little, from the recomputed published 0.0976 to 0.0928. Gains decrease as published scores increase (rank correlation -0.671), without compressing the leaderboard. The fixed cell reverses nine of a hundred and five pairs, versus eleven for calibration on the same fifteen backbones (Appendix H).

Reference reproduction and the Phikon check. Each task recomputes both reference arms before reporting a fixed-cell score. Both must reproduce stored backbone results within 5×10^{-4}. Across 135 cells, the largest deviation is 1.9×10^{-4}. Phikon has stored per-task results for all nine tasks but a published-readout result for only one, because that arm was omitted on the others. The eight missing results were recomputed using the same implementation as all other stored rows. This reproduces their per-task arms bit for bit, with zero deviation, and each recovered published-readout value matches its published row within 0.1%.

L.1 PROVENANCE OF THE PUBLISHED PENALTY

Classification-probe origin of the penalty. The benchmark fixes the ridge penalty at one hundred divided by feature dimension and gene count, overriding the supplied penalty. The same expression, with class count replacing gene count, is the coefficient $1/C$ in UNI's released classification probe (Chen et al., 2024), where $C = dk/100$ for k classes. The objectives differ. The classification probe multiplies half the unsquared weight norm by $1/C$ and adds it to sample-averaged cross-entropy. Ridge adds α times the squared norm to squared error summed over training rows. The same constant is therefore weaker per sample by the training-row count. We compare only this scaling factor. Across Table B.1's $1{,}293$ to $67{,}651$ training spots, it accounts for 3.1 to 4.8 of the 6.1 orders between published and cross-validated penalties, leaving 1.3 to 3.0. This establishes implemented behaviour, not the authors' intent.

Evidence in the benchmark appendix. HEST's appendix compares projected ridge, raw-embedding ridge and gradient-boosted trees. Raw ridge reorders the encoders. ResNet50 scores 0.2842 and Phikon 0.2974, above H-optimus-0 at 0.2754, Virchow2 at 0.2762 and GigaPath (Xu et al., 2024) at 0.2652. Trees keep the published order at both ends, with H-optimus-0 first at 0.3956 and ResNet50 last at 0.326. The authors write that *"[d]irectly applying Ridge regression may unfairly penalize models with larger embedding dimensions"* and *"[t]o guarantee a fairer and more objective comparison, we chose to utilize PCA reduction"*. This pattern is consistent with a penalty set too low, which the authors attribute to width and address through projection. A projection regularises implicitly, by an amount that differs per encoder, and at the published penalty it supplies almost all of the regularisation the readout applies (Appendix L.2).

L.2 EFFECTIVE DEGREES OF FREEDOM AT THE PUBLISHED, FIXED AND SELECTED CELLS

Each calibration run stores the 512 leading squared singular values of its standardised outer-training design, allowing exact evaluation of Equation 1 without refitting. Table L.2 summarises 189 encoder–task–fold fits across seven pre-registered encoders. Squared singular values scale with training-spot count, which ranges from $1{,}293$ to $67{,}651$. At the published penalty, every fit retains more than 0.99998 of its 256 degrees of freedom. At the fixed cell, retained fraction tracks training size (Spearman $\rho = 0.94$ over 189 fits), ranging from median 0.12 on SKCM to 0.78 on CCRCC. The same penalty thus shrinks the smallest task several times more strongly. Inner-selected penalties increase with task size ($\rho = 0.73$ between training spots and selected α, and 0.87 across nine task medians). Selected retained fractions are only weakly related to size ($\rho = -0.21$). Dimension 512 is selected in 157 of 189 fits. In 162 fits, selected degrees of freedom are below the published readout's 256, with median 140 and interquartile range 78 to 197.

Transferring a retained fraction instead of a penalty. The spectral pattern suggests transferring df/d rather than α. The rule was fixed before computation with knowledge of the preceding results. This is therefore a post hoc description, not a registered test. For each scored task, the target is median df/d at H-Optimus-1's selected cells on the other eight tasks, using the same folds as the task-excluded constant. Each scored fold fixes $d = 512$ and selects the grid α whose df/d on its own training spectrum is nearest that target. Only training-side quantities enter. Across seven pre-registered encoders, the rule gains $+0.0285$ over the published readout, versus $+0.0240$ for the task-excluded constant, $+0.0287$ for the all-task constant and $+0.0309$ for per-task selection. It exceeds the task-excluded constant on five of nine tasks and is lower on three, so the advantage is not uniform. Appendix Q reports a registered out-of-sample test on the held-out ST-Net cohort.

M THE TRAINED COMPARATOR, RERUN WITH AND WITHOUT TEST-FOLD SELECTION

Summary. The released-procedure rerun is within 0.005 of the published ten-task mean. CALIBRE exceeds it by 0.0234 on nine tasks and 0.0211 on ten. With the released fixed schedule and one final evaluation, the comparator scores 0.3837 and 0.3534, reflecting overfitting on some tasks. Also selecting learning rates on test folds from 2.5×10^{-4}, 5×10^{-4} and 10^{-3} at one seed gives a nine-task upper bound of 0.4462. CALIBRE remains higher on eight of nine tasks ($p = 0.0137$). Among the nine tasks favouring CALIBRE in the main rerun comparison, HCC's 0.0008 margin lies

Table L.2: Retained fraction df/d from Equation 1, by task in increasing training-size order. Values are medians across seven pre-registered encoders and outer folds, except the published column, which reports minima. s^2_{256} is the 256th squared singular value of the standardised training design. The fixed cell uses $d = 512$, $\alpha = 10^4$. Selected columns use inner-fold choices.

task	folds	training spots	s^2_{256}	published	fixed cell	selected	selected α
SKCM	2	1,517	578	0.99999	0.12	0.43	10^3
LUNG	2	2,603	1,124	0.99999	0.18	0.71	10^3
READ	2	4,204	1,562	0.99999	0.23	0.26	10^4
PAAD	3	5,561	2,091	1.00000	0.26	0.37	10^4
COAD	2	7,826	3,260	1.00000	0.33	0.33	10^4
LYMPH_IDC	4	14,972	8,422	1.00000	0.50	0.10	10^5
IDC	4	29,651	12,207	1.00000	0.58	0.56	10^4
PRAD	2	31,355	14,181	1.00000	0.61	0.19	10^5
CCRCC	6	61,468	35,076	1.00000	0.78	0.31	10^5

within the 0.0051 seed range, so that count is descriptive. The protocol and complete task results follow.

Table 1 reports the trained model's published scores. Here, identical inputs are used for two reruns, one following the released procedure and one omitting test-fold-based selection, both compared with the published row.

Test-fold epoch selection in the released procedure. The released procedure evaluates each outer fold's held-out samples after every epoch, retains the epoch with the highest held-out Pearson correlation, and stops after twenty evaluations without improvement. The reported fold score is thus a maximum over test-set evaluations, a known source of optimistic bias (Varma & Simon, 2006; Cawley & Talbot, 2010). This behaviour is documented in the released source and a public repository issue, but not the paper. Epoch-wise maximisation introduces upward selection bias whose magnitude the paper does not bound.

Two arms, identical except for that step. The first arm follows the released procedure with only the two compatibility changes below. The second removes per-epoch evaluation, selection and early stopping. It trains for the released 100 epochs, saves checkpoints every 10 epochs and evaluates test samples once at the end. Model, loss, sampler, optimiser, hyperparameters and metric remain identical. Training-slide epoch selection was excluded before execution because seven of twenty-nine outer folds have only one training slide. Over the epochs it evaluates, the released arm's test-selected maximum is an upper bound on any training-selected epoch. Learning rate and all other settings retain published defaults without a search. Both arms use the same frozen UNI features (Chen et al., 2024), extracted by the benchmark procedure. Spot alignment passes a check that rejects a deliberately misassigned copy. Each arm uses three seeds, matching the paper, with the decision rule registered before execution.

Compatibility changes applied to both arms. Two compatibility changes are required to run the release. An unsupported model-construction argument is removed because the release implements neither the argument nor its named operation. The module otherwise remains as defined. Per-gene correlations use double precision, matching older library behaviour. This avoids an incompatibility between single-precision values from the numerical libraries used here and the result-recording procedure. The precision change affects scores by 2×10^{-8}. Both changes precede all results and apply identically to both arms.

Machine assignment. Seven tasks use the first machine. CCRCC, PRAD and IDC, the most computationally demanding, use a second machine requiring a newer numerical-library version for its GPU. Inputs are byte-identical across machines, and both arms of each task share a machine. Replicating one cell across machines changes the score by approximately the seed variability.

Table M.1: Trained-comparator reruns on identical inputs, with three seeds per arm and seed ranges. The released procedure selects epochs on test folds. The fixed schedule trains for 100 epochs and is evaluated once.

task	published	as released	range	fixed schedule	range	CALIBRE	before smoothing
IDC	0.587	0.5884	0.0067	0.5752	0.0041	0.6252	0.6243
PRAD	0.421	0.4125	0.0111	0.2277	0.1071	0.4400	0.4397
PAAD	0.507	0.5045	0.0040	0.4674	0.0079	0.5300	0.5300
SKCM	0.704	0.7026	0.0026	0.6531	0.0099	0.7171	0.7150
COAD	0.326	0.3187	0.0051	0.2904	0.0288	0.3642	0.3654
READ	0.240	0.2327	0.0067	0.2076	0.0165	0.2642	0.2633
CCRCC	0.332	0.3176	0.0121	0.2304	0.0181	0.3013	0.2963
LUNG	0.610	0.6099	0.0053	0.5774	0.0175	0.6332	0.6323
LYMPH_IDC	0.305	0.3005	0.0040	0.2241	0.0202	0.3225	0.3105
HCC	0.124	0.1146	0.0051	0.0806	0.0071	0.1154	0.1166
nine tasks	0.448	0.4430	0.0029	0.3837	0.0175	0.4664	0.4641
ten tasks	0.415	0.4102	0.0027	0.3534	0.0162	0.4313	0.4293

Reproduction criterion and the CCRCC exception. The registered reproduction criterion requires a ten-task mean within 0.010 of published values and each task within three published standard deviations or 0.010, whichever is larger. The released-procedure rerun scores 0.4102 over ten tasks (Table M.1). Nine reproduce, but CCRCC scores 0.3176 versus the published 0.332. It is therefore reported as a rerun of the released procedure, not a reproduction of that published result. All CCRCC cells use the second machine. Repeating its first seed on the first machine gives 0.3241, above all second-machine seeds and within 0.008 of the published value. Hardware may explain part of the gap, but one repeated configuration cannot distinguish its effect from seed variation.

Against the rerun as released. With STFlow retaining its test-selected epochs, CALIBRE leads by 0.0234 on nine tasks and 0.0211 on ten, above every seed mean. It is higher on eight of nine and nine of ten tasks (one-sided task-level Wilcoxon $p = 0.0059$ and 0.0049). Fold counts, 23 of 27 and 24 of 29, are descriptive only because folds are nested within tasks. The sole losing task is CCRCC, by 0.0163.

Test-selected learning-rate bound. The primary rerun fixes the released learning rate at 5×10^{-4}. A design registered before execution also evaluates 2.5×10^{-4} and 10^{-3} at the first seed. Each task's bound takes the best of all three rates and, for CCRCC, the first-machine replica. Because both rate and epoch are selected on test folds, this bounds training-side selection over the same grid from above. The additional rates use the first machine, whereas three released-rate tasks use the second. Taking a maximum means a machine offset can only raise the bound. Table M.2 reports first-seed task means.

The best common rate is 10^{-3}. CALIBRE exceeds it by 0.0220 over nine tasks, positive on eight ($p = 0.0059$). Against task-specific rate bounds, the margin is 0.0202 over nine tasks, positive on eight ($p = 0.0137$), and 0.0179 over ten, where HCC joins CCRCC as a loss. The released-rate rerun remains the registered comparison.

Fixed-schedule comparison, as registered. The released fixed schedule with one final evaluation scores 0.3837 and 0.3534. The registered superiority rule requires CALIBRE to exceed the seed mean by more than both the seed range and the noise floor. Margins of 0.0827 and 0.0779 exceed thresholds of 0.0175 and 0.0162. CALIBRE is higher on 25 of 27 and 27 of 29 folds.

Contribution of epoch selection to the gap. Rescoring saved checkpoints separates schedule length from epoch selection. Choosing the best checkpoint on each test fold recovers most of the gap between arms, attributing it mainly to selection rather than training. The released 100-epoch schedule exceeds the useful range on some tasks. On one PRAD fold, checkpoint scores vary by 0.2 despite smoothly falling training loss. The conclusion therefore does not rely on the fixed schedule. Choosing each task's most favourable epoch with access to test folds gives 0.4197 over nine tasks and 0.3881 over ten, below CALIBRE on all ten tasks.

Table M.2: The trained comparator at three learning rates, with epochs selected on test folds, beside CALIBRE. All values are first-seed task means. The bound takes each task's best rate and, for CCRCC, the first-machine replica.

task	CALIBRE	2.5×10^{-4}	5×10^{-4}	10^{-3}	bound
IDC	0.6252	0.5807	0.5855	0.5906	0.5906
PRAD	0.4400	0.4054	0.4086	0.4149	0.4149
PAAD	0.5300	0.5004	0.5029	0.5034	0.5034
SKCM	0.7171	0.6946	0.7033	0.6955	0.7033
COAD	0.3642	0.3170	0.3208	0.3268	0.3268
READ	0.2642	0.2318	0.2334	0.2362	0.2362
CCRCC	0.3013	0.3229	0.3100	0.3200	0.3241
LUNG	0.6332	0.6096	0.6123	0.6081	0.6123
LYMPH_IDC	0.3225	0.2914	0.2981	0.3045	0.3045
HCC	0.1154	0.1149	0.1154	0.1181	0.1181
nine tasks	0.4664	0.4393	0.4417	0.4444	0.4462
ten tasks	0.4313	0.4069	0.4091	0.4118	0.4134

Compute cost. CALIBRE is not faster in wall-clock time. Its seven calibrated readouts took 6.1 hours across ten tasks on a shared CPU server, with three concurrent fits. Released-procedure comparator training took 2.0 GPU-hours for one seed, 6.3 for three seeds and 11.3 for the three fixed-schedule seeds. Per task, one released run required 0.17 to 1.30 times the readout time of CALIBRE. Both timings start from cached features. Encoding adds to both, especially CALIBRE, which processes every patch through seven encoders rather than one. For 13,620 HER2ST spots, sequential encoding on one RTX A6000 took 6,546 seconds, or 1.8 hours, versus 291 seconds for the comparator's encoder. Each timing includes a published-readout fit and therefore slightly overstates encoding time. Smoothing adds one kernel per slide and no fit because inner predictions are already computed during selection. CALIBRE avoids predictor training, not wall-clock cost.

N PATIENT-HELD-OUT BREAST COHORT

The HEST experiments cover the benchmark's ten tasks. CALIBRE is also applied unchanged to a previously unevaluated cohort on a different platform at coarser resolution. The trained-model comparison follows Appendix M.

Cohort. HER2ST comprises spatially profiled HER2-positive breast tumours (Andersson et al., 2021), with 36 sections from eight patients. Four patients contribute six sections each and four contribute three. Its 13,620 spots have 100 µm diameter on a 200 µm grid, compared with 55 µm spots on a 100 µm grid in all 72 benchmark samples. We use the HEST parent-dataset release (Jaume et al., 2024), whose patch and expression processing matches the benchmark. None of the benchmark's 72 samples belongs to these sections or their study.

Cohort construction under the benchmark's task rules. The released task-construction procedure specifies how to add a task, although the benchmark paper does not. It selects the fifty most variable genes after a per-sample expression filter and groups folds by study and patient. We reproduced both steps exactly, recovering the ordered gene lists and fold partitions for all ten existing tasks. Applied to this cohort, the same rules give eight folds, each holding out all sections from one patient. The gene panel was fixed before any model was fitted to or evaluated on this cohort.

Patient-grouped inner selection. The seven encoders, their weights, target and both readouts remain unchanged. The calibrated readout selects its projection size and penalty on inner folds of the training data, which on the benchmark are grouped by slide because no patient identifier is released. Here, each patient contributes up to six sections. Holding out a single section would leave related sections from the same tumour in training and provide an optimistic selection criterion. Inner folds therefore hold out whole training patients. Section-wise inner folds, the default procedure, are retained as a pre-registered sensitivity arm.

Table N.1: Results for all eight held-out patients. Scores average per-gene Pearson correlations pooled across each patient's spots. Frozen readouts use seed 1. Trained-model rows average three seeds within patients. Table P.1 reports dense and matched coarse training-side selectors based on separate inner trajectories. UNI2-h is included as the best of seven encoders on this cohort, selected using test-patient outcomes. The best fixed epoch is likewise test-selected.

patient	A	B	C	D	E	F	G	H	mean
sections	6	6	6	6	3	3	3	3	
CALIBRE	.142	.629	.476	.263	.030	.035	.276	.264	0.2644
before smoothing	.125	.597	.446	.242	.039	.033	.274	.255	0.2515
before smoothing, inner folds by section	.128	.610	.434	.208	.018	.033	.268	.268	0.2458
seven uncalibrated, averaged	.123	.607	.421	.193	.011	.029	.261	.262	0.2385
H-Optimus-1, calibrated	.131	.569	.445	.305	.055	.034	.258	.220	0.2521
UNI2-h, calibrated	.134	.585	.459	.329	.069	.037	.283	.244	0.2675
trained, epoch chosen on test	.115	.626	.437	.349	.033	.023	.309	.302	0.2742
trained, fixed schedule	.055	.576	.275	.051	.025	.007	.188	.245	0.1778
trained, best fixed epoch on test	.091	.607	.391	.259	.006	.016	.293	.298	0.2451
trained, coarse training-side selector	.091	.607	.391	.123	.006	.016	.293	.298	0.2281

Decision rule, fixed before scoring. For each held-out patient, the contrast is CALIBRE minus the comparator. A descriptive win for CALIBRE requires the eight-patient mean to exceed the registered margin of 0.0019 and at least six differences to be positive. The primary comparator is fixed-schedule STFlow. Secondary comparators are the uncalibrated seven-encoder average and the pre-named single encoder. These comparisons also change smoothing because only CALIBRE receives it. Appendix O matches smoothing access. Test-patient epoch selection is reported separately and disclosed as such.

Registered verdicts. Table N.1 reports every patient. Against the primary fixed-schedule comparator, CALIBRE gains a mean 0.0866 and is higher on all eight patients, meeting the registered rule. It exceeds the uncalibrated seven-encoder average by 0.0260 on all eight, meeting the descriptive rule, but also adds smoothing. Appendix O isolates calibration under matched smoothing. Individual calibration improves all seven encoder means by $+0.010$ to $+0.048$. Relative to the calibrated H-Optimus-1 readout, the full method gains 0.0124 on six of eight patients and meets the rule. Before smoothing, the ensemble trails by 0.0005 and is higher on four of eight. On this cohort, the ensemble exceeds its named member only after smoothing. Because this comparison changes both averaging and smoothing, it does not isolate an averaging benefit. Against the test-selected best calibrated encoder, UNI2-h, it is lower by 0.0031 and higher on four of eight. The unsmoothed ensemble is lower by 0.0160, higher on two of eight and loses under the registered rule. Against STFlow with test-patient-selected epochs, it is lower by 0.0098 and higher on four of eight patients, failing the win rule. Grouping inner folds by section exposes sister tumour sections during selection. This lowers the unsmoothed ensemble from 0.2515 to 0.2458 on held-out patients and lowers six of seven encoders. The section-grouped arm still exceeds the fixed schedule by 0.0680 on seven of eight patients.

The fixed schedule overtrains on this cohort. Rescoring saved fixed-schedule checkpoints on held-out patients, as in Appendix M, gives a three-seed mean of 0.2451 at epoch 10, falling to 0.1781 at epoch 100. Table N.1's 0.1778 instead uses the random state left by training, without a seed reset. Six patients peak at the first checkpoint under all three seeds. Released early stopping occurs at epochs 22 to 48, with test-selected epochs from 2 to 28. Selection precedes the first checkpoint for 19 of 24 patient–seed pairs. This explains why the released test-epoch arm, 0.2742, exceeds the test-selected best checkpoint per patient, 0.2490. Schedule choice therefore accounts for almost all the primary margin. Against the best fixed epoch selected on test patients, CALIBRE gains 0.0193 on six of eight. Conclusions do not rely on the primary fixed-schedule verdict.

A coarse training-side selector. Neither of the two initial registered arms fairly estimates trained-model performance on this cohort. Training-side selection was unavailable on seven of twenty-nine benchmark folds, but every external-cohort fold contains seven training patients. A third arm was

Table O.1: Matched readout arms on three cohorts, mean per-gene Pearson correlation. HEST averages three seeds within each of nine tasks. HER2ST (eight patients) and ST-Net (nineteen patients, cohort-wide gene panel) use seed one with patients held out. The single encoder is H-Optimus-1. Its published HEST row is our reproduction (0.4228 versus the leaderboard's 0.4229). Each arm selects P-ES separately on training data. Rows with the same encoder count compare calibration, and rows with the same readout compare averaging (Appendices O and Q).

Encoders	Readout	HEST		HER2ST		ST-Net	
		no P-ES	P-ES	no P-ES	P-ES	no P-ES	P-ES
One	published	0.4228	0.4377	0.2167	0.2408	0.0788	0.0943
One	calibrated	0.4465	0.4543	0.2521	**0.2697**	0.1024	0.1120
Seven	published	0.4524	0.4575	0.2385	0.2511	0.0932	0.1038
Seven	calibrated	0.4641	**0.4664**	0.2515	0.2644	0.1120	**0.1208**

registered after observing the test-selected arm and two fixed-schedule seeds, but before rescoring any cohort checkpoint. For each outer patient, seven inner models train on six patients and validate on the seventh, using the released procedure for 100 epochs with checkpoints every 10. The epoch maximising the mean inner score, with ties favouring earlier epochs, is selected from the outer fixed-schedule checkpoints for each of three seeds. No model was fitted to or evaluated on the held-out patient before its epoch was fixed. Selection uses one seed on the second machine in Appendix M. Only the chosen epoch is transferred to the first machine. The loss rule requires CALIBRE's mean difference to be below -0.0019, with at most two positive patients.

The selector chooses epoch 10 for seven patients and epoch 20 for D. Evaluating the fixed-schedule checkpoints at those epochs gives STFlow 0.2281. CALIBRE leads by a mean 0.0363, on six of eight patients, meeting the descriptive rule. Under a fair coin, at least six positives among eight has probability 0.145, so this is a registered decision rule, not a fixed-level significance test. The paired patient interval, $[-0.0103, +0.0830]$, includes zero. Before smoothing, the ensemble leads by 0.0234 on five patients, one short of the rule, making the verdict depend on smoothing. Patient D contributes substantially. Its inner score rises from 0.2009 at epoch 10 to 0.2078 at 20. At epoch 20, STFlow scores 0.1226 on D versus CALIBRE's 0.2627. Excluding D leaves a 0.0215 mean difference, positive for five of seven patients. Seven patients select the earliest candidate, while test selection favours earlier epochs on 19 of 24 pairs. The dense diagnostic in Appendix P was designed in response to this observation. The coarse comparison meets the descriptive rule with an interval spanning zero, whereas the dense comparison does not. This selector and the matched coarse selector of that appendix differ because the latter uses separate inner trajectories and selects epoch 40 for D, rather than 20.

Equal-weight averaging on this cohort. On HEST, equal averaging exceeds the best member selected per fold with hindsight by mean 0.0033, but is lower on eight of twenty-nine folds. Before smoothing on this cohort, it trails H-Optimus-1 on four of eight patients and the best member chosen separately per patient with hindsight on seven of eight. Within folds, the rank correlation between prediction variance across spots and encoder score averages $+0.137$ on HEST and -0.504 here. High-variance predictions therefore belong to stronger HEST encoders but weaker external-cohort encoders, suggesting that equal averaging may favour weaker members here. This explanation, its HEST-fold test and falsifier were formulated after the patient results were known. The relationship between that correlation and averaging loss has the predicted direction but is not significant (Spearman -0.329, $p = 0.081$), leaving the explanation unconfirmed. Rescaling members before averaging is untested and would use these patient outcomes for selection. CALIBRE therefore remains unchanged. Matched smoothing gives the single encoder a higher mean than the ensemble (Appendix O), without establishing an ensemble-specific smoothing benefit. Patients E and F remain near zero under every frozen or trained arm.

O MATCHED COMPONENT CONTROLS WITH SMOOTHING HELD AVAILABLE

A full-recipe comparison cannot attribute the gain to calibration or averaging when only the full method has smoothing. Six registered controls give the baselines the same smoothing access as

Table O.2: Matched-control mean Pearson correlations. The two published single-encoder rows are descriptive additions outside the registered four-test family, as detailed below. HEST averages three seeds within tasks. HER2ST averages patients at seed 1. Seed-one HEST values elsewhere, including the uncalibrated average's 0.4520, differ from these three-seed means.

Readout	HEST, 9 tasks	HER2ST, 8 patients
Published single encoder	0.422792	0.216725
Published single encoder + P-ES	0.437692	0.240816
Calibrated single encoder	0.446531	0.252052
Calibrated single encoder + P-ES	0.454253	0.269650
Uncalibrated ensemble	0.452418	0.238454
Uncalibrated ensemble + P-ES	0.457534	0.251053
Calibrated ensemble	0.464063	0.251512
Calibrated ensemble + P-ES	0.466376	0.264422

Table O.3: Four matched contrasts, with tasks or patients as units, not seeds, spots or genes. Paired bootstrap mean intervals are unadjusted for multiplicity.

Cohort	Contrast	Mean difference	95% mean interval	Positive	Exact p	Holm p
HEST	calibration	$+0.0088$	$[+0.0006, +0.0164]$	$7/9$	0.0977	0.3906
HEST	averaging	$+0.0121$	$[-0.0003, +0.0227]$	$8/9$	0.0977	0.3906
HER2ST	calibration	$+0.0134$	$[+0.0011, +0.0288]$	$6/8$	0.1094	0.3906
HER2ST	averaging	-0.0052	$[-0.0289, +0.0117]$	$5/8$	0.8438	0.8438

CALIBRE to assess those components. They comprise calibrated H-Optimus-1, the uncalibrated seven-encoder average and the calibrated seven-encoder average, each with and without P-ES. Smoothed arms share the fixed parameter grid but select their own parameters on training data. This matches smoothing access without imposing identical settings on different predictions. Encoders, splits, genes and metric remain fixed. The calibrated ensemble reproduces every stored fold score of Appendix V within 10^{-9}, with identical smoothing selections.

HEST includes all nine published tasks and seeds $1, 2, 3$, averaged within tasks before comparison. HER2ST includes all eight patients at seed 1, reusing HER2ST's existing seed-one results unchanged. All 28 task–seed or cohort cells are retained. Both cohorts informed development before this diagnostic was registered, so neither is an untouched validation cohort. Table O.2 gives the six registered arms and two published single-encoder arms added after that registration. Contrasts condition on the other analysis choices. They neither separate penalty from projection nor establish that the components add or interact.

Prespecified contrasts and inference. Calibration compares smoothed calibrated and uncalibrated ensembles. Averaging compares the smoothed calibrated ensemble with the smoothed calibrated single encoder. Positive differences favour the full method. Two-sided exact paired Wilcoxon tests use average ranks for ties and omit exact zeros, followed by Holm correction over all four contrasts. No contrast has an exact-zero difference. Unadjusted paired percentile mean intervals use 20,000 bootstrap draws with seed 20260919. Resampling uses tasks for HEST and patients for HER2ST, sharing indices across each cohort's two contrasts. Seeds are not independent replicates. Mean intervals do not replace corrected rank tests.

All four contrasts appear in Table O.3. Calibration has positive means in both cohorts and unadjusted mean intervals excluding zero. None of the four rank tests survives Holm correction at 0.05. Averaging has a positive HEST mean increment and a negative HER2ST increment, with both intervals including zero. On HER2ST, the smoothed single encoder scores 0.269650 versus the full method's 0.264422. This does not establish a reversal, but CALIBRE's advantage over an unsmoothed single encoder cannot be attributed to averaging alone. Non-significance does not establish the absence of a component effect.

The published single encoder under the same protocol. The six registered arms omit the published single-encoder readout, leaving its calibration and smoothing effects inseparable. An addi-

Table O.4: H-Optimus-1 single-encoder contrasts. Units are HEST tasks after seed averaging or HER2ST patients. Intervals are unadjusted paired bootstrap mean intervals. p denotes the two-sided exact paired Wilcoxon test. Interaction is calibration with P-ES minus calibration without P-ES.

Cohort	Contrast	Mean difference	95% mean interval	Positive	Exact p
HEST	calibration, both smoothed	$+0.0166$	$[+0.0064, +0.0265]$	$8/9$	0.0195
HEST	calibration, neither smoothed	$+0.0237$	$[+0.0131, +0.0338]$	$8/9$	0.0078
HEST	P-ES on the published readout	$+0.0149$	$[+0.0096, +0.0210]$	$9/9$	0.0039
HEST	P-ES on the calibrated readout	$+0.0077$	$[+0.0046, +0.0111]$	$9/9$	0.0039
HEST	interaction	-0.0072	$[-0.0101, -0.0046]$	$0/9$	0.0039
HER2ST	calibration, both smoothed	$+0.0288$	$[+0.0063, +0.0544]$	$5/8$	0.1094
HER2ST	calibration, neither smoothed	$+0.0353$	$[+0.0136, +0.0587]$	$7/8$	0.0234
HER2ST	P-ES on the published readout	$+0.0241$	$[+0.0094, +0.0403]$	$7/8$	0.0234
HER2ST	P-ES on the calibrated readout	$+0.0176$	$[+0.0053, +0.0307]$	$7/8$	0.0234
HER2ST	interaction	-0.0065	$[-0.0117, -0.0013]$	$2/8$	0.0781

tional H-Optimus-1 experiment evaluates those cells with identical folds, seeds, genes and P-ES grid. Each of 28 task–seed or cohort cells first reproduces the calibrated single-encoder result, requiring agreement in folds, patients, smoothing choices, per-gene scores and task scores. The published aggregate agrees with its stored value within 10^{-16}. The analysis was specified with results from only 15 of 28 cells already available. Its primary contrast is calibration with matched smoothing, using tasks or patients, the same exact Wilcoxon test and paired bootstrap intervals ($20{,}000$ draws, seed 20260924). Because both cohorts informed development, these contrasts remain descriptive and outside the four-test Holm family.

Table O.4 reports all ten contrasts. With P-ES available to both readouts, calibration raises H-Optimus-1 from 0.437692 to 0.454253 on HEST, improving eight of nine tasks. On HER2ST it raises the score from 0.240816 to 0.269650, improving five of eight patients. The HER2ST mean interval excludes zero, but its rank test does not reach 0.05. Without smoothing, gains are larger at $+0.0237$ and $+0.0353$. P-ES improves both readouts, more strongly for the published one. The calibration and P-ES effects partly overlap, with a negative interaction on every HEST task. On H-Optimus-1, calibration retains a positive increment under matched smoothing, smaller than the unsmoothed comparison suggests.

P Denser training-side epoch selection for the external comparator

The coarse selector in Appendix N chooses its earliest checkpoint for seven patients, motivating a diagnostic with earlier candidates. Registration used knowledge of those cohort outcomes, so this is not untouched external confirmation. No inner curve or score from the coarse selector is reused. Both selectors were fixed before outer evaluation. For each of eight outer patients, seven seed-1 inner models train on six patients and validate on the seventh. Each of the 56 inner runs follows one 100-epoch trajectory. The dense grid is $\{1, \ldots, 20, 30, 40, \ldots, 100\}$, comprising 28 scored epochs. The matched coarse grid is $\{10, 20, \ldots, 100\}$ on the same trajectories. Both maximise the equally weighted mean over seven inner patients, breaking ties towards earlier epochs. A training-only control on the same machine exactly reproduces the first scored trajectory's weights at epochs 10 and 20.

For patients A–H, frozen dense epochs are $(4, 6, 8, 3, 3, 2, 4, 3)$ and matched coarse epochs are $(10, 10, 10, 40, 10, 10, 10, 10)$. Appendix N's coarse selector chooses epoch 20 for D and remains a separate comparison. Inner fitting uses an NVIDIA L20. Outer evaluation uses an NVIDIA RTX A6000 with the original training procedure, retaining all three seeds, eight patient folds and 100-epoch trajectories. Added scoring preserves random-number state and model mode. All 48 checkpoint comparisons show exactly identical model weights at epochs 10 and 20. Only the two frozen epochs per fold enter this diagnostic. CALIBRE retains its unchanged seed-one result.

Patient-paired results. Table P.2 reports paired patient differences. CALIBRE exceeds dense-selected STFlow by mean 0.0022, positive for five of eight patients, with an interval spanning zero.

Table P.1: All eight held-out patients. STFlow averages three seeds within each patient. CALIBRE retains its seed-one predictions. Dense and coarse selectors use the common inner trajectories described here, not the coarse selector in Appendix N. No patient or seed is excluded.

Patient	Dense epoch	Coarse epoch	CALIBRE	Dense STFlow	Coarse STFlow	CALIBRE − dense
A	4	10	0.1417	0.1073	0.0909	+0.0343
B	6	10	0.6288	0.6210	0.6072	+0.0078
C	8	10	0.4762	0.4157	0.3910	+0.0605
D	3	40	0.2627	0.3498	0.1018	−0.0871
E	3	10	0.0305	−0.0031	0.0056	+0.0336
F	2	10	0.0351	0.0133	0.0163	+0.0218
G	4	10	0.2759	0.3040	0.2928	−0.0281
H	3	10	0.2644	0.2893	0.2984	−0.0249
Mean	–	–	0.2644	0.2622	0.2255	+0.0022

Table P.2: Differences paired across eight patients after averaging STFlow seeds within patients. Intervals are unadjusted paired t intervals with seven degrees of freedom. Twenty-four patient–seed pairs do not constitute twenty-four independent patients.

Paired comparison	Mean difference	Patient 95% interval	Positive
CALIBRE − dense STFlow	+0.0022	$[-0.0370, +0.0415]$	$5/8$
CALIBRE − coarse STFlow	+0.0389	$[-0.0125, +0.0903]$	$6/8$
Dense − coarse STFlow	+0.0367	$[-0.0355, +0.1088]$	$5/8$

The registered descriptive rule requires a mean above 0.0019 and at least six positive patients for a win, or below -0.0019 with at most two positives for a loss. Neither direction is met against the dense selector. The coarse comparison meets the descriptive win rule, but its interval also includes zero. This rule is not a significance test. Dense minus coarse averages 0.0367, again with an interval spanning zero. D contributes the largest gain, 0.2480. Patients E, F and H have negative dense-minus-coarse differences and remain included. No patient exclusion, threshold or test was added.

Superiority on HER2ST remains unresolved under the stronger training-side selector, establishing neither equivalence nor CALIBRE's inferiority. The diagnostic changes neither the HEST experiment nor the CALIBRE recipe and does not test reduced annotation budgets. Its design followed prior HER2ST outcomes, so it is not untouched external confirmation.

Q A HELD-OUT PATIENT COHORT

Cohort. The breast cancer cohort of He et al. (2020), processed in HEST-1k (Jaume et al., 2024), contains 68 Spatial Transcriptomics sections from 23 patients. Clinical identifiers link four patients to HER2ST's clinical patients E, G, H and I. Patient I has no released section. All four are excluded. The remaining 56 sections comprise 19 patients and 23,726 spots, all invasive ductal carcinoma, imaged at $20\times$. Spots have $100\,\mu$m diameter on a $200\,\mu$m grid, with a CC BY 4.0 licence. None of these sections belongs to HEST-Benchmark or HER2ST, and none had previously been scored in this study. The clinical crosswalk establishes patient separation from HER2ST, but institution and platform are shared. The analysis plan was registered before any section was encoded.

Construction. Construction follows the benchmark rules reproduced exactly on all released HEST-Benchmark tasks before HER2ST construction. Fifty genes are selected by cohort-wide variance. Holding out each patient with all sections gives 19 outer folds. Gene selection uses expression, not predictions, but includes held-out patients' expression, as in the benchmark. A registration made after the cohort-panel analysis repeats the primary contrast, selecting fifty genes by the same rule from each fold's eighteen training patients alone. These panels overlap the cohort panel by 26 to 45 genes (median 40). Calibration raises H-Optimus-1 from 0.0811 to 0.1033, a gain of $+0.0222$ (paired bootstrap interval $[+0.0103, +0.0339]$), positive for 16 of 19 patients (two-sided exact Wilcoxon $p = 0.0039$). The cohort-panel gain is $+0.0235$ with $p = 0.0024$. The same implementation exactly reproduces two registered cohort-panel fold scores. One fold-panel gene is absent

Table Q.1: Unsmoothed retained-fraction transfer on ST-Net, averaged over 19 patients. Recovery is the rule's gain over the published readout divided by the per-fold calibration gain. The final column gives the selected penalty, identical across folds.

Encoder	published	fixed cell	rule	calibrated	recovered	rule α
H-Optimus-1	0.0788	0.0833	0.0940	0.1024	65%	10^5
H-optimus-0	0.0783	0.0844	0.0946	0.1027	67%	10^5
UNI2-h	0.0715	0.0794	0.0904	0.1101	49%	10^5
Virchow2	0.0788	0.0814	0.0910	0.1119	37%	10^5
Midnight-12k	0.0825	0.0910	0.0910	0.0955	65%	10^4
Phikon-v2	0.0728	0.0840	0.0840	0.1077	32%	10^4
Phikon	0.0786	0.0853	0.0853	0.1011	30%	10^4

from a patient's three held-out sections. It is treated as undetected, giving an undefined correlation and excluding it from that patient's score. Readout and P-ES selection both use inner folds holding out whole training patients.

Arms and statistics. Each of seven encoders receives published, calibrated and fixed-cell readouts. The fixed cell in Section 4 was chosen on HEST before this cohort was used and requires no selection here. Each arm is evaluated with and without P-ES, selected on inner patient folds and applied within sections. H-Optimus-1 is the single encoder named in the HEST registration. Ensembles average each readout over all seven encoders. Scores average per-gene Pearson correlations across each held-out patient's spots, with patients as independent units. The primary contrast is unsmoothed calibrated minus published H-Optimus-1, tested by two-sided exact paired Wilcoxon over 2^{19} sign assignments. Paired percentile bootstrap intervals use 20,000 draws. Five secondary contrasts are Holm-adjusted within their family. The initial registration excludes trained comparators. A separate registration made afterwards adds the model described under *The trained comparator* below.

Results. Table Q.2 reports all arms and Table Q.3 all registered contrasts. The primary result meets the registered criterion. Calibration raises H-Optimus-1 from 0.0788 to 0.1024, improving 16 of 19 patients (paired bootstrap interval $[+0.0106, +0.0364]$, $p = 0.0024$). All seven encoders improve by $+0.0130$ to $+0.0385$ (Table Q.4). With P-ES available to both readouts, calibration gains $+0.0177$ for the single encoder and $+0.0170$ for the ensemble. Averaging adds $+0.0088$ over the smoothed calibrated single encoder, improving 12 of 19 patients, but its interval reaches zero. Its contribution is not established on this cohort either. The HEST fixed cell improves all encoders by only $+0.0026$ to $+0.0112$. Inner-selected penalties range from 10^5 to 10^8, with encoder modes between 10^6 and 10^8. These are two to four decades above the fixed cell and mostly above HEST selections, consistent with larger optimal penalties at lower signal-to-noise ratios. Mean Pearson is near 0.1, versus about 0.45 on HEST. UNI2-h selects the grid maximum in every fold, so its optimum may lie beyond the search. As α increases, predictions approach α^{-1} times a linear-kernel weighted sum of the training expression profiles. Pearson correlation ignores that scale, placing the grid-edge readout near a similarity-weighted training average. CALIBRE selects mixing weight 0.7 on 18 of 19 folds. Of 23,726 tissue spots, 23,680 have patches and are scored.

Transferring the retained fraction. Part A of a second registration tests Appendix L.2's post-hoc rule out of sample. It was written before either quantity was computed on this cohort. The target, $\tau = 0.3708$, is fixed from HEST as median df/d at H-Optimus-1's inner-selected cells across 27 folds. For every ST-Net outer fold and encoder, $d = 512$. The grid penalty α is chosen to place df/d on that fold's training spectrum closest to τ, using no test-side quantity. At $\alpha = 10^4$, this implementation reproduces every registered fixed-cell patient score within 10^{-8}. Table Q.1 reports the results. The rule gains $+0.0152$ for H-Optimus-1, improving 17 of 19 patients ($p = 0.0006$). It recovers 65% of the per-fold gain, versus 19% for the fixed cell. Every encoder improves over its published readout ($p \leq 0.0062$), but recovery is only 30% to 67%, and three encoders select the fixed-cell penalty. This meets the registered interpretation of partial transfer. Retained fraction transfers better than penalty, but optimal shrinkage still varies by cohort.

Table Q.2: All ST-Net arms, averaged over 19 held-out patients using per-gene Pearson correlation at seed 1. P-ES is selected on inner patient folds and applied separately within each section.

Encoders	Readout	no P-ES	P-ES
H-Optimus-1	published	0.0788	0.0943
H-Optimus-1	calibrated	0.1024	0.1120
H-Optimus-1	fixed cell	0.0833	0.0985
Seven encoders	published	0.0932	0.1038
Seven encoders	calibrated	0.1120	**0.1208**
Seven encoders	fixed cell	0.0968	0.1067

Table Q.3: Registered ST-Net contrasts with patients as units ($n = 19$). Intervals are paired percentile bootstrap intervals for the mean (20,000 draws). p denotes two-sided exact paired Wilcoxon over 2^{19} sign assignments. Holm correction covers the five secondary contrasts. The primary contrast and S1 are unsmoothed. Pos. counts patients with positive differences.

	Contrast	Mean	95% interval	Pos.	p	Holm p
Primary	one encoder, calibrated − published	$+0.0235$	$[+0.0106, +0.0364]$	16/19	0.0024	–
S1	one encoder, fixed cell − published	$+0.0045$	$[+0.0016, +0.0076]$	15/19	0.0108	0.0433
S2	as primary, both with P-ES	$+0.0177$	$[+0.0047, +0.0306]$	15/19	0.0141	0.0433
S3	CALIBRE − seven published with P-ES	$+0.0170$	$[+0.0032, +0.0314]$	15/19	0.0124	0.0433
S4	CALIBRE − one calibrated with P-ES	$+0.0088$	$[-0.0002, +0.0197]$	12/19	0.1447	0.1447
S5	CALIBRE − one published, unsmoothed	$+0.0419$	$[+0.0241, +0.0617]$	18/19	0.00034	0.0017

Encoder ordering. A post hoc analysis comparing the ordering of the seven encoders under the two unsmoothed readouts was specified and fixed before computation. The published readout ranks Midnight-12k first and UNI2-h last. Calibration ranks Virchow2 first, UNI2-h second and Midnight-12k last, with Kendall τ of -0.52 between rankings ($p = 0.14$, seven encoders). Sixteen of 21 pairs reverse. For H-optimus-0 versus UNI2-h, UNI2-h versus Midnight-12k, and Midnight-12k versus Phikon-v2, paired patient-bootstrap intervals exclude zero under both readouts. These three reversals are resolved by that criterion, but none of sixteen survives Holm correction across 42 paired Wilcoxon tests. Shared patients make the encoder comparisons dependent.

The trained comparator. Part B was registered before any model training and evaluates STFlow with its released procedure and UNI features. Within each outer fold, eighteen training patients are sorted by identifier and allocated cyclically to three groups. Each inner model trains on two groups and scores the third at epochs 1–20 and every tenth epoch through 100. The epoch with the highest mean inner score is recorded and fixed before outer training. Each outer model then trains to that epoch under three seeds and is scored once on the held-out patient. Selected epochs range from 1 to 70 (median 10). A 20-epoch training-only replica of the first inner model matches the scored model exactly at epochs 10 and 20, confirming that scoring preserves its trajectory. Descriptive contrasts against STFlow on UNI give $+0.0329$ for smoothed calibrated H-Optimus-1 (13 of 19 patients, $p = 0.049$) and -0.0003 for published H-Optimus-1 ($p = 0.89$). A further registration repeats the procedure on H-Optimus-1 using four RTX A4000 GPUs. Epochs range from 1 to 80 (median 5), with eight of nineteen folds selecting the first epoch because subsequent inner curves are flat and noisy. Appendix R compares both models with the readouts.

Prior outcome exposure at registration. The initial ST-Net registration used an unscored cohort. Each subsequent registration preceded its own experiments but was written with knowledge of preceding results on the same patients. Table Q.5 identifies that dependence for every registration. The kidney cohort (Appendix T) was registered before any of its data were downloaded, so its results carry no such dependence.

R THE CALIBRATED READOUT ON THE TRAINED MODEL'S OWN ENCODER

Design. STFlow uses UNI (Chen et al., 2024). Every rerun receives the same cached UNI features with alignment verified as in Appendix M. Before any calibrated UNI readout was fitted, a

Table Q.4: Unsmoothed encoder scores on ST-Net, averaged over 19 patients. The final column gives the modal selected penalty and its fold count. Dimensions are 512 throughout, except one H-optimus-0 fold selecting 64.

Encoder	published	calibrated	fixed cell	calibration gain	fixed-cell gain	modal penalty
H-Optimus-1	0.0788	0.1024	0.0833	+0.0235	+0.0045	10^6, 19/19
H-optimus-0	0.0783	0.1027	0.0844	+0.0244	+0.0061	10^6, 18/19
UNI2-h	0.0715	0.1101	0.0794	+0.0385	+0.0079	10^8, 19/19
Virchow2	0.0788	0.1119	0.0814	+0.0332	+0.0026	10^7, 19/19
Midnight-12k	0.0825	0.0955	0.0910	+0.0130	+0.0085	10^6, 16/19
Phikon-v2	0.0728	0.1077	0.0840	+0.0349	+0.0112	10^6, 18/19
Phikon	0.0786	0.1011	0.0853	+0.0225	+0.0067	10^7, 10/19

Table Q.5: Prior results available at each ST-Net registration, ordered by registration. Each plan preceded its specified experiments, but subsequent plans had access to results on the same cohort.

	analyses registered	results on this cohort already known
1	calibration of H-Optimus-1 (primary), seven encoders, CALIBRE against the published ensemble, cohort-wide panel	none
2	retained-fraction transfer; STFlow on UNI with epochs chosen on training patients, cohort-wide panel	the readouts of 1
3	the calibrated readout on UNI, STFlow's input	the readouts of 1
4	the primary of 1 with gene panels chosen in each training fold	the readouts of 1
5	STFlow on H-Optimus-1 features	the readouts of 1
6	the three trained-model comparisons of Table 2, with fold-internal panels and P-ES for both predictors	1 to 5, including the cohort-wide STFlow comparisons
7	STFlow with a learning-rate and epoch budget chosen on training patients	1 to 6

registration specified CALIBRE's unchanged single-encoder procedure on all three cohorts at seed one. Published and calibrated readouts are scored with and without inner-training-fold-selected P-ES. The published readout's predictions use a fixed thread count. Relative to four threads, individual predictions differ by at most 0.025, while the published nine-task mean changes by 1.1×10^{-5}. That mean is 0.3857, versus the leaderboard's 0.3873. The discrepancy comes from IDC (0.5735 versus 0.589), with the other eight tasks agreeing within 0.0006. Every calibrated prediction reproduces its stored counterpart within 10^{-8}.

HEST. Table R.1 reports every HEST task. Calibration raises UNI from 0.3857 to 0.4171 over nine published tasks, and smoothing raises it to 0.4236. Against the nine-task means averaged from STFlow's published frozen UNI and trained task scores, 0.3773 and 0.4480, this closes 65% of the gap with encoder input unchanged. The calibrated, smoothed readout trails the test-epoch-selected rerun's 0.4430 on eight of nine tasks, but exceeds the fixed-schedule rerun's 0.3837 on all nine. No test is reported because the former selects on test folds.

HER2ST. With patients held out, calibrated and smoothed UNI scores 0.2733, versus 0.2622 for densely training-selected STFlow. The mean difference is +0.0111, positive for five of eight patients (paired t interval $[-0.0175, +0.0397]$, two-sided exact Wilcoxon $p = 0.55$). Neither registered direction is met. Without smoothing, the difference is −0.0007 ($p = 0.95$). The single calibrated UNI readout exceeds CALIBRE's seven-encoder score of 0.2644 on this cohort.

Encoder-by-predictor comparison on the held-out cohort. Table R.2 crosses UNI and H-Optimus-1 with calibration and STFlow on 19 held-out patients using a cohort-wide gene panel. All other choices use training patients, and only readouts receive smoothing. STFlow scores 0.0792 versus its own encoder's calibrated, smoothed readout at 0.1082. The readout is higher for 14 of

Table R.1: UNI, STFlow's input encoder, on HEST at seed one, using mean per-gene Pearson correlation. Trained-model columns are the three-seed reruns in Appendix M. One selects each fold's epoch on test data as released. The other uses the released fixed schedule.

	published readout		calibrated readout		trained model	
task	no P-ES	P-ES	no P-ES	P-ES	test-fold epoch	fixed schedule
IDC	0.5735	0.5826	0.5780	0.5837	0.5884	0.5752
PRAD	0.2944	0.3085	0.3845	0.3860	0.4125	0.2277
PAAD	0.4806	0.4915	0.4903	0.4951	0.5045	0.4674
SKCM	0.6345	0.6538	0.6588	0.6700	0.7026	0.6531
COAD	0.2620	0.2692	0.3026	0.3065	0.3187	0.2904
READ	0.1836	0.1917	0.2109	0.2130	0.2327	0.2076
CCRCC	0.2399	0.2572	0.2444	0.2533	0.3176	0.2304
LUNG	0.5464	0.5589	0.5856	0.5899	0.6099	0.5774
LYMPH_IDC	0.2560	0.2796	0.2984	0.3145	0.3005	0.2241
HCC	0.0784	0.0840	0.1048	0.1061	0.1146	0.0806
nine published tasks	0.3857	0.3992	0.4171	0.4236	0.4430	0.3837
ten tasks	0.3549	0.3677	0.3858	0.3918	0.4102	0.3534

Table R.2: Encoders crossed with predictors on ST-Net with cohort-wide gene panels, averaged over 19 held-out patients. Readouts use registered seed-one predictions. STFlow averages three seeds at inner-patient-selected epochs. Contrasts report mean differences, paired t intervals, positive-patient counts and two-sided exact Wilcoxon p. Matched rows compare predictors using the same encoder. Each of the first three rows is primary in its own registration. The seven-encoder CALIBRE scores 0.1208.

encoder	published	calibrated	calibrated, P-ES	STFlow
UNI	0.0798	0.0992	0.1082	0.0792
H-Optimus-1	0.0788	0.1024	0.1120	0.0900
contrast	mean	95% interval	positive	p
CALIBRE − STFlow, UNI	+0.0416	[+0.0126, +0.0707]	14/19	0.0071
UNI, calibrated, P-ES − STFlow	+0.0291	[+0.0018, +0.0564]	14/19	0.0258
H-Optimus-1, calibrated, P-ES − STFlow	+0.0220	[−0.0079, +0.0519]	13/19	0.17
UNI, calibrated − STFlow	+0.0200	[−0.0033, +0.0433]	11/19	0.10
H-Optimus-1, calibrated − STFlow	+0.0124	[−0.0100, +0.0347]	12/19	0.28
CALIBRE − STFlow, H-Optimus-1	+0.0308	[+0.0017, +0.0599]	14/19	0.0204
STFlow, H-Optimus-1 − UNI	+0.0108	[−0.0045, +0.0261]	11/19	0.26
calibrated, P-ES, H-Optimus-1 − UNI	+0.0038	[−0.0144, +0.0220]	11/19	0.68

19 patients ($p = 0.026$). On H-Optimus-1, STFlow rises to 0.0900, and the matched +0.0220 difference is unresolved ($p = 0.17$). Neither unsmoothed matched difference is resolved. Holm correction across three registered trained-model comparisons retains CALIBRE versus STFlow on UNI (adjusted $p = 0.021$), while the matched UNI result becomes 0.052. Encoder changes are unresolved under both STFlow and the calibrated readout.

Fold-internal panels and matched smoothing. A second registration, written after the preceding comparisons but before evaluating any of its own cells, repeats three trained-model comparisons with two changes. Each fold selects fifty genes from its eighteen training patients using the benchmark rule. STFlow also receives the frozen readouts' smoother, with bandwidth and mixing weight selected on inner patient folds at its frozen epoch and applied to every outer seed within held-out sections. One panel gene is absent from all three held-out sections of one patient and is counted as undetected, so its correlation is undefined and it drops out of that patient's score in every arm. STFlow epochs are selected from 57 inner runs per encoder on training patients only and frozen before any outer model was trained. Three reproduction checks pass before comparison. Frozen readouts exactly reproduce two earlier cohort-panel folds. A cohort-panel STFlow outer fit reproduces its stored score within 10^{-16}. Two training runs are tensor-equal at epochs 10 and 20. Primary contrasts are CALIBRE minus STFlow on UNI, and calibrated readout minus STFlow on UNI and H-

Table R.3: Secondary contrasts of the fold-internal ST-Net comparison in Table 2, in the same format, without multiplicity correction.

contrast	mean	95% interval	positive	p
UNI, neither smoothed	$+0.0218$	$[-0.0027, +0.0462]$	$13/19$	0.0955
H-Optimus-1, neither smoothed	$+0.0180$	$[-0.0050, +0.0409]$	$13/19$	0.12
STFlow, UNI, P-ES − none	$+0.0060$	$[+0.0003, +0.0118]$	$12/19$	0.0874
STFlow, H-Optimus-1, P-ES − none	$+0.0062$	$[+0.0013, +0.0111]$	$15/19$	0.0071

Optimus-1 with both predictors smoothed. Each uses two-sided exact Wilcoxon across 19 patients, with Holm correction over three contrasts.

Results with fold-internal panels. Table 2 reports all arms. CALIBRE scores 0.1213 versus STFlow on UNI at 0.0765, improving 17 of 19 patients (two-sided exact Wilcoxon $p = 0.00064$, Holm-adjusted 0.0019). This meets the registered rule, and omitting any patient leaves p no greater than 0.0013. With smoothing, STFlow scores 0.0825 on UNI and 0.0915 on H-Optimus-1. Matched calibrated readouts score 0.1075 and 0.1136, improving 13 of 19 patients for each encoder ($p = 0.0955$ and 0.13, both Holm-adjusted 0.19). Neither matched comparison is resolved. CALIBRE remains ahead of smoothed STFlow by $+0.0388$ ($p = 0.0082$) and $+0.0297$ ($p = 0.0361$) on the two encoders, respectively, both secondary comparisons. Frozen scores change little from the cohort-panel results in Table R.2. CALIBRE scores 0.1213 versus 0.1208, while calibrated, smoothed UNI and H-Optimus-1 score 0.1075 and 0.1136 versus 0.1082 and 0.1120. Table R.3 lists the remaining secondary contrasts.

Descriptive controls. Two descriptive controls, without a decision rule, were added to that registration before any readout evaluation was complete. Both use the same panels and patients. Within-section permutation preserves feature distributions while removing spot-level image information, lowering calibrated H-Optimus-1 from 0.1033 to 0.0031. The large-penalty limit, α^{-1} times a linear-kernel weighted sum of training expression, scores 0.0938 on H-Optimus-1 and 0.1095 on UNI2-h. UNI2-h selects the grid maximum 10^8 in every fold. Its calibrated score of 0.1095 coincides with the limit, indicating a grid-bounded rather than interior optimum on this cohort.

S A TRAINING-ONLY TUNING BUDGET FOR THE TRAINED MODEL

Design. In the fold-internal ST-Net comparison, STFlow selects only its epoch on training patients, whereas each frozen readout searches 44 cells. A further registration, written before its experiments began, gives STFlow on UNI a joint budget of three learning rates, 2.5×10^{-4}, the released 5×10^{-4} and 10^{-3}, crossed with its 28 scored epochs. The released training code uses Adam without weight decay or a schedule, so the learning rate is its main optimisation setting. Each outer fold takes the configuration with the highest mean over its three inner patient groups at seed one, with ties going to the released rate and then to the earlier epoch, and the choices were fixed before any outer model was trained. Outer folds run at three seeds, and P-ES for the tuned model is selected on the inner predictions at the chosen configuration. The two primary contrasts, CALIBRE minus tuned STFlow and the smoothed calibrated UNI readout minus smoothed tuned STFlow, are Holm-corrected together. Before these experiments, the modified training implementation exactly reproduced a stored outer-fold result on the same GPU, both with an explicit learning rate and with its default. STFlow on H-Optimus-1 keeps epoch-only selection.

Results. Fourteen folds chose 2.5×10^{-4}, four chose 10^{-3} and one kept the released rate, with epochs from 1 to 40. The fold that kept the released rate reproduces the stored default result exactly. Tuning lowered STFlow on held-out patients, from 0.0765 to 0.0580 without smoothing (4 of 19 patients improved, $p = 0.060$) and from 0.0825 to 0.0600 with P-ES (3 of 19, $p = 0.0385$). Both registered primary contrasts favour the readout. CALIBRE leads the tuned model by $+0.0633$ for 17 of 19 patients ($p = 0.00042$, Holm-adjusted 0.00084), and by $+0.0613$ when the tuned model has P-ES ($p = 0.00064$). The smoothed calibrated UNI readout leads the smoothed tuned model by $+0.0475$ for 15 of 19 ($p = 0.011$, Holm-adjusted 0.011). Dropping any one patient keeps both primaries' means positive. Because the budget lowered the comparator, the shared-encoder

result does not show that the readout exceeds a well-tuned trained model, and against the released configuration the same contrast remains unresolved ($p = 0.0955$, Table 2). On this cohort, a learning rate selected on inner folds of eighteen patients did not transfer to held-out patients.

T A SECOND HELD-OUT COHORT ON A DIFFERENT PLATFORM

Cohort. The human kidney atlas of Lake et al. (2023), processed in HEST-1k (Jaume et al., 2024), contains 23 Visium sections from 22 individuals, comprising reference tissue and biopsies with acute kidney injury or diabetic kidney disease. Spots have 55 µm diameter on a 100 µm grid, imaged at 20×, with a CC BY 4.0 licence. The cohort differs from HER2ST and the ST-Net cohort in institution, platform, organ and disease. None of its sections had been scored in this study. The analysis plan, including the wording used for its outcome in the abstract, was registered before any section was downloaded. Only sample counts, patient counts, spot counts and disease labels from the HEST-1k metadata had been read.

Construction and arms. The protocol is that of the fold-internal ST-Net comparison (Appendix R) with the cohort changed. Holding out each patient with all sections gives 22 outer folds. Each fold's fifty genes are chosen by the benchmark's variance rule on its 21 training patients. Readout and P-ES selection use inner folds that hold out one training patient. STFlow receives UNI or H-Optimus-1 features, its epochs are chosen on three inner patient groups at seed one and fixed before any outer model is trained, and outer folds are run at three seeds. Before any readout or model was scored on this cohort, the same code run on ST-Net reproduced the stored gene panel, the stored readouts of two folds for UNI and H-Optimus-1 within 10^{-8}, the stored STFlow datasets file by file, the epoch choice, the aggregation of outer-fold scores, and one stored STFlow outer cell on the same card within 10^{-6}.

Tests. The first primary contrast is unsmoothed calibrated minus published H-Optimus-1. The other three compare CALIBRE with STFlow on UNI, and the smoothed calibrated readout with the smoothed trained model on UNI and on H-Optimus-1, with Holm correction over the three. Each is a two-sided exact paired Wilcoxon over 2^{22} sign assignments, with patients as units, a paired t interval and the count of patients improved. Two rules fixed on other data, the HEST fixed cell and the retained fraction $\mathrm{df}/d = 0.37$, are also applied without selection.

Results. Table T.1 reports both encoders and every registered contrast. The primary calibration result meets its criterion. Calibration raises H-Optimus-1 from 0.1880 to 0.1991 for all 22 patients ($p = 4.8 \times 10^{-7}$), and all seven encoders improve, by $+0.0077$ to $+0.0127$ on 18 to 22 patients each. With P-ES on both sides the gain is $+0.0099$, again for all 22 patients. CALIBRE exceeds STFlow on UNI by $+0.0349$ for 20 of 22 patients ($p = 9.1 \times 10^{-6}$, Holm-adjusted 2.7×10^{-5}), meeting the registered rule, and by $+0.0302$ when STFlow receives the same smoother ($p = 0.00018$). With a shared encoder and smoother, the readout leads by $+0.0124$ on UNI ($p = 0.033$) and $+0.0102$ on H-Optimus-1 ($p = 0.050$), and both Holm-adjusted values are 0.066. Dropping any one patient keeps every primary's mean positive. STFlow epochs frozen from the inner curves lie between 40 and 100. Averaging adds $+0.0097$ over the smoothed calibrated single encoder for 17 of 22 patients ($p = 0.00026$), whereas calibration inside the ensemble adds $+0.0012$ over the smoothed published ensemble for 13 of 22 ($p = 0.32$), so on this cohort the ensemble gains from averaging rather than from calibration.

Transfer of the locked rules. Inner selection returns the HEST cell, $d = 512$ and $\alpha = 10^4$, in 131 of 154 encoder folds, and UNI2-h selects $\alpha = 10^5$ in 20 of its 22. The HEST cell applied unchanged recovers 96 to 121% of the per-fold calibration gain across the seven encoders, and the retained-fraction rule 84 to 118%. On H-Optimus-1, the rule gains $+0.0093$, improving 18 of 22 patients ($p = 0.00094$). Correlations near 0.2 are about twice those of the ST-Net cohort, and the selected penalties match HEST's rather than ST-Net's 10^6 to 10^8, which is consistent with the rules transferring here and not there.

Descriptive controls. The two descriptive controls of the ST-Net comparison, applied to the same kidney panels, give 0.0003 for calibrated H-Optimus-1 with held-out embeddings permuted within

Table T.1: The kidney cohort with gene panels chosen inside each training fold, mean over 22 held-out patients. P-ES parameters and STFlow epochs are selected on inner patient folds. STFlow averages three seeds. Each + P-ES column adds P-ES to the arm on its left. Contrasts give the mean difference, paired t interval, count of patients improved and two-sided exact Wilcoxon p. The first contrast is the registered calibration primary, and the next three are the trained-model primaries, Holm-corrected together. The seven-encoder CALIBRE scores 0.2144.

encoder	published	calibrated	+ P-ES	STFlow	+ P-ES
UNI	0.1833	0.1920	0.1967	0.1795	0.1843
H-Optimus-1	0.1880	0.1991	0.2047	0.1889	0.1945

contrast	mean	95% interval	positive	p	Holm p
Calibration, H-Optimus-1	$+0.0111$	$[+0.0087, +0.0135]$	$22/22$	4.8×10^{-7}	
CALIBRE − STFlow, UNI	$+0.0349$	$[+0.0227, +0.0472]$	$20/22$	9.1×10^{-6}	2.7×10^{-5}
UNI, both with P-ES	$+0.0124$	$[+0.0011, +0.0237]$	$15/22$	0.0329	0.0658
H-Optimus-1, both with P-ES	$+0.0102$	$[-0.0024, +0.0228]$	$17/22$	0.0501	0.0658
UNI, neither smoothed	$+0.0125$	$[+0.0014, +0.0236]$	$16/22$	0.0190	
H-Optimus-1, neither smoothed	$+0.0102$	$[-0.0012, +0.0217]$	$16/22$	0.0501	
STFlow, UNI, P-ES − none	$+0.0047$	$[+0.0011, +0.0084]$	$17/22$	0.0141	
STFlow, H-Optimus-1, P-ES − none	$+0.0057$	$[+0.0025, +0.0088]$	$19/22$	0.0013	
CALIBRE − STFlow + P-ES, H-Optimus-1	$+0.0199$	$[+0.0073, +0.0326]$	$18/22$	0.0017	
CALIBRE − STFlow + P-ES, UNI	$+0.0302$	$[+0.0168, +0.0435]$	$18/22$	0.00018	

Table U.1: Both methods lose accuracy with fewer annotated patients, with a larger loss for CALIBRE. Means cover eight held-out patients and three subset draws, using seed 1 for both methods at every budget. STFlow uses Appendix N's coarse training-side selector. Its seven-patient score is seed 1 of the arm whose three-seed mean in Appendix N is 0.2281. The dense selector in Appendix P was not tested at reduced budgets.

annotated patients	trained	CALIBRE	margin	CALIBRE ahead on	one-sided p
two	0.1810	0.1974	$+0.0164$	$6/8$	0.074
four	0.2085	0.2541	$+0.0456$	$8/8$	0.0039
seven	0.2260	0.2644	$+0.0384$	$6/8$	0.055

each section, against 0.1991 unpermuted. The large-penalty limit scores 0.1175 on H-Optimus-1 and 0.1338 on UNI2-h, below their calibrated 0.1991 and 0.2058. Unlike on ST-Net, every selected penalty for both encoders lies inside the grid.

U ANNOTATION-BUDGET SENSITIVITY ON HER2ST

A frozen readout selects only a penalty and a projection before fitting a linear map, which suggests that it should degrade less than a model trained end to end when annotated tissue is removed. This prediction was registered before measurement and is not supported on this cohort.

Design. Annotation budget k is the total number of labelled training patients available to an arm, including model selection. Budgets are $k \in \{2, 4, 7\}$ on the external cohort, where each outer fold has seven training patients. Some HEST folds contain only one training slide and cannot support this curve. For each held-out patient and $k < 7$, three distinct size-k subsets are drawn using the registered seed and shared between arms. STFlow selects an epoch within every budget as in Appendix N. It uses k leave-one-patient-out fits, selects the maximum mean inner-curve epoch, then trains on the full subset to that epoch. This requires $k + 1$ fits per draw and 192 in total. CALIBRE retains all three steps, fitting on the subset and selecting (d, α) per encoder and (σ, c) by internal leave-one-patient-out validation. This requires a minimum budget of two. Recomputing the full-budget arm from Appendix N gives exactly zero maximum patient-score deviation, confirming that patient restriction is the only change.

Table V.1: Task means at seed one before smoothing and after P-ES with its parameters selected on inner folds. The registered decision rule covers the nine tasks, so HCC is listed separately.

task	before smoothing	CALIBRE	gain
IDC	0.6243	0.6252	+0.0009
PRAD	0.4397	0.4400	+0.0002
PAAD	0.5300	0.5300	−0.0000
SKCM	0.7150	0.7171	+0.0021
COAD	0.3654	0.3642	−0.0012
READ	0.2633	0.2642	+0.0009
CCRCC	0.2963	0.3013	+0.0050
LUNG	0.6323	0.6332	+0.0009
LYMPH_IDC	0.3105	0.3225	+0.0120
nine tasks	0.4641	0.4664	+0.0023
HCC, outside the rule	0.1166	0.1154	−0.0012

Registered decision. From seven patients to two (Table U.1), STFlow loses 0.0450 and CALIBRE loses 0.0671. The difference is -0.0221, with CALIBRE losing less on three of eight patients (one-sided Wilcoxon $p = 0.875$ for the favourable direction). This meets the registered unfavourable branch and does not support greater annotation-budget robustness.

Budget-specific results. Against the coarse-selector comparator, CALIBRE remains ahead at every budget, but its advantage narrows from $+0.0384$ to $+0.0164$. The prediction that this advantage grows as annotation decreases fails. STFlow's flatter curve partly reflects its low full-budget score of 0.2260, leaving less room to decline. On patients D and F, STFlow scores higher with two training patients than with seven, by 0.0640 and 0.0028. These increases are consistent with epoch-selection sensitivity, but do not isolate its contribution. Smoothing gains $+0.0129$ with seven patients, $+0.0119$ with four and -0.0004 with two. At the smallest budget it helps twelve of twenty-four cells, versus twenty of twenty-four at four patients. Its parameters use inner validation within the subset, giving only two folds at the minimum budget. Smoothing is retained, and every cell selects a non-zero weight.

The unfavourable branch. The favourable branch requires at least six of eight patients to favour CALIBRE. The registration imposes no patient-count condition on the unfavourable branch, and none is applied.

V OUTPUT SMOOTHING ON CALIBRE, WITH ITS PARAMETERS SELECTED ON INNER FOLDS

The extension that passes in Table W.1 was evaluated on a readout outside CALIBRE. Its parameters used task-excluded selection, contrary to the registered inner-fold procedure. The calibrated-ensemble evaluation instead selects both parameters on inner folds, with its design and decision rule registered before execution. Its results support smoothing as CALIBRE's third step.

Construction. For each outer fold, every encoder's calibrated readout is refitted at its selected cell within that fold's inner splits. Equal averaging gives the ensemble's inner predictions. P-ES (Rolf et al., 2020) is applied within each inner held-out slide, using bandwidth $\sigma \in \{0.5, 1, 1.5, 2, 3, 5, 8\}$ in median nearest-neighbour spacings and mixing weight $c \in \{0, 0.1, \ldots, 1\}$. The pair maximising the readout's inner criterion is applied to held-out slides. Ties favour smaller c, leaving CALIBRE unchanged under a flat criterion. Only predictions and spot coordinates are used from held-out slides.

Reproduction checks. Recomputed outer predictions must match those underlying CALIBRE's reported scores. Refitted inner scores at selected cells must also match the selection records. Across twenty-eight task–seed evaluations, maximum deviations are 0 and 1.1×10^{-16}, respectively.

Smoothing gains and destructive controls. Smoothing raises the nine-task mean from 0.4641 to 0.4664 (Table V.1), gaining $+0.0023$ at each of three seeds. Seven of nine tasks improve (one-sided task-level Wilcoxon $p = 0.037$), chiefly LYMPH_IDC at $+0.0120$ and CCRCC at $+0.0050$, while COAD loses 0.0012. Both registered controls are negative and below the required quarter-gain threshold. Shrinkage towards the held-out slide's mean prediction at the same weight loses 0.0063, and coordinate permutation loses 0.0050. Neither shrinkage nor kernel weights alone explains the gain. A separately evaluated lattice smoother uses identical selection and gains $+0.0022$, but is excluded from all decision rules. All twenty-seven folds select non-zero weights, ranging from 0.1 to 0.9 (median 0.4). Bandwidth reaches the grid minimum on fourteen folds and maximum on eight.

On HER2ST. With unchanged patient-grouped inner selection, smoothing raises CALIBRE from 0.2515 to 0.2644. It exceeds the unsmoothed uncalibrated average by 0.0260 on all eight patients and the unsmoothed calibrated single encoder by 0.0124 on six of eight. These comparisons change smoothing as well as the named component. With matched smoothing, the single encoder scores 0.2697 (Appendix O). STFlow scores 0.2281 with coarse training-side selection and 0.2622 with dense selection. Against the latter, CALIBRE gains 0.0022 on five of eight patients, with an interval spanning zero (Appendix P). Test-selected epochs remain ahead at 0.2742.

Computational cost. The ten-task evaluation at one seed took two processor-hours on a shared server, almost entirely for refitting. Deployment does not require these refits because selection computes the same inner predictions at every grid cell. The evaluated procedure, however, discards them. Smoothing itself adds one kernel per slide.

W EXCLUDED VARIANTS AND NEGATIVE RESULTS

This section collects the variants that did not enter CALIBRE, namely eight registered extensions that failed their rules, a backbone excluded by the reproduction gate, a transductive normalisation and the pooling component. Every negative result is kept, summarised rather than reported step by step.

Registered extensions. Each extension was registered before measurement, with a mechanism-destroying control or parameter-free alternative as its comparator. Three used the excluded pooling readout from Appendix W.3, as marked in Table W.1. Three add capacity to a calibrated single encoder by shrinking gene-specific cells towards a shared curve, appending spatial context or mapping predictions across genes. None meets its criterion. Two change shrinkage shape through a component-index penalty profile or a rank constraint on gene coefficients. They gain $+0.0007$ and $+0.0012$ on five of seven encoders, below the registered 0.0019 margin. A block-specific penalty for concatenated encoders contributes -0.0040 in its nested removal ablation, positive on two of nine tasks ($p = 0.9727$), and remains 0.0080 below equal averaging. Per-gene encoder routing finds cross-fold agreement of 0.236 versus a permuted null of 0.170, but routing selected on other folds loses 0.0156. Selecting one encoder per gene discards averaging's variance reduction. Shared weights fitted to CALIBRE predictions gain $+0.0016$ ($p = 0.1797$), below the margin and the parameter-free alternative of dropping the two weakest encoders. That alternative does not separate from all seven either. The ninth extension passes on its evaluated readout, but selects two parameters with tasks held out rather than the registered inner folds. Repeating it on the calibrated ensemble with inner-fold selection meets the same rule and gains $+0.0023$ at each of three seeds. It is CALIBRE's smoothing step (Appendix V). On the evaluated readout, lattice smoothing of each encoder's predictions gains $+0.0058$, versus $+0.0067$ for tuned post-estimation smoothing (Rolf et al., 2020). The operation is established. The contribution is measuring its role in this recipe.

Table W.1 lists each extension, control or comparator, evaluated readout and registered result. Each registration fixes its grid, selection procedure and control.

W.1 A BACKBONE EXCLUDED BY THE REPRODUCTION GATE

A fourth backbone was evaluated as additional confirmation under three pre-registered predictions. The first was a reproduction gate. The unmodified readout had to match the published nine-task

Table W.1: All extensions, with mean gains over each row's comparator, positive encoders or tasks and unadjusted one-sided task-level tests. Only output smoothing passes its thresholds on the listed readout, although its parameters use task-excluded rather than registered inner-fold selection. It also passes when repeated on the calibrated ensemble with the registered selection, becoming the third step (Appendix V). The slide-wide control is an identity on five single-test-slide tasks, prompting a graded control after its results were observed. On CALIBRE, the unfitted average of the top five encoders by published rank exceeds learned shared weights.

extension	what it changes	control or comparator	measured on	result
per-gene cells	each gene's (d, α), shrunk toward the all-gene choice	cells reassigned to permuted genes	calibrated readout	-0.0006, $3/7$ encoders, $p = 0.63$
spatial context	embedding augmented with the mean of six nearest spots	six random spots of the slide	calibrated readout	-0.0017, $3/7$, $p = 0.79$
cross-gene map	predictions mapped across genes	rows permuted in fit and use	calibrated readout	-0.0336, $0/7$, $p = 0.99$
penalty profile	penalty as a function of component index	profile on shuffled components	calibrated readout	$+0.0007$, $5/7$, $p = 0.37$
rank constraint	rank of the gene coefficients	rank after rotating the gene axes	calibrated readout	$+0.0012$, $5/7$, $p = 0.37$
per-block penalty	one penalty per encoder in a concatenation	registered bar, one encoder with the weight	bar with the pooling weight	$+0.0044$ over the bar, $p = 0.29$, and 0.0080 below equal averaging
output smoothing	each prediction mixed with its nearest spots'	slide-wide mean, rewired neighbours	readout with the pooling weight	$+0.0058$, $7/7$, $p = 0.0020$, and tuned post-estimation smoothing $+0.0067$
per-gene router	one encoder per gene, chosen on other folds	equal weights	average with the pooling weight	-0.0156
shared weights	one weight vector per task, fit on the other eight	equal weights	CALIBRE	$+0.0016$, $7/9$ tasks, $p = 0.18$

mean within 1%. Failure would invalidate the backbone's configuration and all downstream comparisons.

The gate failed on all six completed tasks. The backbone exceeds its published row by a mean 4.31%, with deviations of $+1.9\%$, $+2.4\%$, $+3.9\%$, $+4.5\%$, $+9.4\%$ and $+9.5\%$. This systematic positive bias indicates a different representation rather than noise. The arm was voided, and the remaining three tasks, the largest, were not run.

The model card and benchmark specify different representations. The card recommends concatenating the class token with the mean patch token for classification, doubling the width. The benchmark's encoder library defaults to the class token alone, with no override in its model configuration. The published row therefore uses half the width evaluated in this arm. Both representations are valid uses of the encoder, but only the class-token representation matches the leaderboard.

As pre-registered, this arm is void rather than a negative result because it evaluates a different representation. Its apparent gains, $+0.033$ to $+0.067$ per task and positive on every task, were the largest observed. They are not valid leaderboard-comparison results because the leaderboard never evaluated that representation.

An independent ablation by PLUTO-4G (Padigela et al., 2025) confirms which representation the benchmark uses. It reports 0.427 for the class token alone and 0.432 for concatenated class and mean patch tokens, but uses the former as its benchmark score.

The reproduction check also identified a second configuration error. Three simultaneous discrepancies caused deviations of 12.8% to 37.4% across three tasks, although none was documented in the papers or raised an error. A separately registered corrected arm retained the three predictions and added a stopping rule. Another gate failure would exclude the backbone rather than permit a third extraction, preventing repeated extraction from becoming a search for agreement.

The corrected extraction. Class-token-only extraction under that registration reproduces the published nine-task mean within 0.01%, with no task deviation above 0.1%. Calibration gains $+0.0344$ (Table A.2). Both remaining registered predictions concern the excluded pooling arm and hold (Appendix W.3).

W.2 THE TRANSDUCTIVE VARIANT

Standardising embeddings using each slide's own statistics jointly uses all spots on that slide. This is not label leakage, but it is not comparable with the twenty-six published encoders evaluated with training-slide statistics only. Its gain is also unstable. One task contributes 91%, another loses, and PRAD loses under two independent backbones. Neither this variant nor the headline recipe uses held-out labels. The headline readout is fitted on training slides only. Its smoother uses only held-out predictions and coordinates, with parameters selected on training slides before held-out targets are scored.

W.3 THE EXCLUDED POOLING COMPONENT

The pooling arm adds one parameter to each calibrated readout. Ridge is fitted to $\mathbf{Y} - \lambda\mathbf{M}$ rather than $\mathbf{Y}$, where $\mathbf{M}$ contains row-aligned slide means computed from training rows. The weight $\lambda \in \{0.0, 0.1, \ldots, 1.0\}$ is selected jointly with d and α by inner cross-validation over training slides using the benchmark metric. No slide mean is restored at prediction, and $\lambda = 0$ recovers calibration alone. Full slide-mean removal is the panel-econometric within transformation (Mundlak, 1978). Partial removal relates to random-effects quasi-demeaning (Balestra & Nerlove, 1966), with batch correction the closest biological analogue (Johnson et al., 2007).

After calibration, the pooling weight's contribution does not separate from zero. On H-optimus-0, pooling gains $+0.0357$ over the published readout, comprising $+0.0302$ from calibrating the fixed hyperparameters and $+0.0055$ from pooling ($+0.00545$ on identical inner folds). Pooling improves four tasks, lowers three and leaves two unchanged (one-sided Wilcoxon $p = 0.1875$, bootstrap interval $[-0.0004, +0.0141]$, which contains zero). Across seven pre-registered backbones, means range from $+0.0023$ to $+0.0078$, all positive. Task-level inference gives $+0.0050$ at $p = 0.36$, while joint sign flips across backbones within tasks give one-sided $p = 0.15$. Excluding CCRCC leaves $+0.0011$. That task's pooling partition depended on an uncontrolled slide-identifier ordering. Replacing each backbone's score by its mean over controlled orderings changes the seven-backbone gain from $+0.00496$ to $+0.00471$. Calibration uses a different partition and returns the same CCRCC score under all twelve hash seeds tested on H-optimus-0. CALIBRE excludes pooling. Table W.1 identifies the three registered extensions evaluated with it.

X DETAILED COMPARISONS WITH PRIOR WORK

X.1 FROZEN-FEATURE BENCHMARKS AND THEIR READOUTS

He et al. (2020) established histology-to-expression prediction. The frozen encoders currently evaluated on that task are general-purpose pathology foundation models (Chen et al., 2024). HEST-Benchmark (Jaume et al., 2024) defines the protocol examined here. HESCAPE (Gindra et al., 2025), the largest independent benchmark in this area, adopts its ridge readout without specifying a penalty or releasing readout code. It also defers slide-level batch effects as an open problem. Appendix C evaluates penalty sensitivity under a fully specified protocol.

X.2 PROBE DESIGN AND REPRESENTATION COMPARISONS

Natural-language probing treats a probe as a modelling choice, not a neutral instrument (Hewitt & Liang, 2019; Voita & Titov, 2020; Pimentel et al., 2020). Penalty cross-validation is routine in vision (Goyal et al., 2019) and brain encoding, including per-target penalties (Dupré la Tour et al., 2022). Neural-encoding benchmarks likewise rank frozen networks through linear mappings (Schrimpf et al., 2018; Conwell et al., 2024). Most frozen-vision studies report little change in relative performance across probes. Kolesnikov et al. (2019) find *"the relative performance of various settings is mostly unchanged"* when replacing a linear probe with a small network. Kornblith et al. (2019) select penalties from forty-five validation candidates. VTAB's reversal concerns frozen readout versus fine-tuning, not two frozen-readout settings (Zhai et al., 2019). The closest positive result is from Lee et al. (2023), who describe probe performance as *"very sensitive to the hyperparameters"* and trace it to unnormalised inputs, which batch normalisation before the linear layer largely removes. Our narrower contribution traces a benchmark's fixed setting to its source and measures its effect on frozen-versus-trained comparisons.

X.3 Spectral dependence of ridge regression

Ridge error depends on representation eigenspectrum, target alignment and penalty (Bordelon et al., 2020; Canatar et al., 2023). The risk comparison of Dhillon et al. (2013) shows that leading-component projection itself regularises. Appendix L.2 measures this established dependence empirically, without proposing a new theory.

X.4 Tuned baselines and evaluation conclusions

Retuning baselines under common protocols has substantially reduced reported advances in metric learning (Musgrave et al., 2020), neural recommendation (Ferrari Dacrema et al., 2019) and language modelling. In the latter, properly regularised long short-term memory networks recovered state-of-the-art performance (Melis et al., 2018). Other work characterises largely irreducible evaluation variance (Bouthillier et al., 2021) or task-dependent rankings (Dehghani et al., 2021). Our result also retunes a published baseline, but identifies a specific cause, a fixed penalty negligible relative to the embedding spectrum. We measure its consequences for encoder rankings and frozen-versus-trained comparisons. The effect is not simply a larger tuning budget. One unchanged cell recovers most of the gain from a published penalty that applies almost no shrinkage (Section 4).

X.5 Trained models and their frozen baselines

Task-trained predictors range from He et al. (2020) through multi-resolution models (Chung et al., 2024) to whole-slide flow matching (Huang et al., 2025a), which reports the strongest published benchmark score. Diffusion models (Zhu et al., 2025) and models that add spatial context to frozen foundation-model features (Nonchev et al., 2025) are further trained alternatives. Their frozen baselines use fixed readouts. For the flow-matching model, this is HEST-1k's random forest with 70 trees. We rerun that model rather than rely solely on its published comparison against an uncalibrated baseline. The results do not establish that trained architectures are unnecessary. STRank (Nishimura et al., 2025) changes the fitting objective to relative-expression ranking, not the frozen readout. A benchmark of eleven trained predictors (Wang et al., 2025) likewise compares architectures rather than readouts. The claim here is that the published benchmark readout conflates representation content with readout choice. Part of the reported trained-model margin therefore belongs to the readout.

X.6 Operations on predictions rather than features

Post-estimation smoothing (Rolf et al., 2020) smooths predictions over a structural index, selecting bandwidth and mixing weight on validation data without test labels. We evaluate it because independent spot-wise prediction discards spatial arrangement. Tuned P-ES outperforms the lattice smoother evaluated here, so the operation is attributed to prior work. Spatial gene-expression smoothing is also established. SPRITE (Sun et al., 2024) post-processes the output of any spatial prediction method over a neighbourhood graph, to impute unmeasured genes rather than predict from histology. We contribute the measurement rather than the operation. CHRep (Wang et al., 2026) calibrates a trained encoder using nearest-neighbour estimates from training slides plus a learned residual. CALIBRE instead selects a frozen readout's penalty and dimension within training folds.

X.7 Encoder-side progress on the same benchmark

MINT (Lee et al., 2026) trains a pathology encoder with a transcriptomic token and scores 0.4401 through the same fixed readout. Encoder training and readout calibration can therefore be combined in principle. Unreleased weights prevent testing that combination. PLUTO-4G (Padigela et al., 2025) scores 0.427 under the same official readout. STPath (Huang et al., 2025b) changes the patient-stratified folds and is not directly comparable.