Supplementary Information

Training-Free Correction of Gene Expression Predictions Using Cancer-Specific and Spatial Priors

Contents

Supplementary Note 1 Ablation of the bio-prior covariance structure

To isolate the functional mechanics of the biological prior, we evaluate three reduced regularizers against the full multi-prior on HEST-bench at the canonical configuration $(\lambda_{\text{bio}}, \lambda_{\text{sp}}) = (0.15, 0.2)$. The first shrinks predictions toward the empirical mean μ_c alone in the spirit of a James-Stein estimator. The second uses a diagonal-only Σ_c^{-1}, which retains each gene's conditional precision but drops all off-diagonal partial correlations. The third is the full Σ_c^{-1} formulation. All variants use the identical correction protocol, isolating the prior's covariance topology as the sole experimental variable.

The mean-shrinkage variant yields $\Delta = 0$ across all baselines because vanilla predictions already lie near the empirical mean. The diagonal-only variant actively degrades performance, systematically underperforming the vanilla baselines by 0.05 to 0.07 in mean PCC, because it rescales each gene independently and pulls co-regulated genes apart. Only the full Σ_c^{-1} moves predictions along the co-expression structure rather than across it. The off-diagonal entries of Σ_c^{-1} are therefore not an incremental refinement but the sole active ingredient driving the efficacy of the bio prior.

Table S1: Reduced-prior ablation on HEST-bench at $(\lambda_{\text{bio}}, \lambda_{\text{sp}}) = (0.15, 0.2)$ (mean PCC across 29 cancer×fold pairs). James-Stein shrinkage toward μ_c alone yields no gain; the diagonal-only Σ_c^{-1} (per-gene scaling without off-diagonal co-expression) underperforms vanilla by 0.05-0.07. The full multi-prior gain therefore comes specifically from the off-diagonal cancer-conditional gene co-expression.

Baseline	Vanilla	μ_c shrinkage (James-Stein)	Diagonal-only Σ_c^{-1}	Full multi-prior
ST-Net	0.2792	0.2792 (+0.0000)	0.2316 (−0.0476)	0.3056 (+0.0264)
GCN	0.3473	0.3473 (+0.0000)	0.2848 (−0.0625)	0.3799 (+0.0326)
HisToGene	0.3620	0.3620 (+0.0000)	0.2900 (−0.0720)	0.3839 (+0.0220)

Supplementary Note 2 Σ_c^{-1} top-partial-correlation pairs

We extract the top-20 partial-correlation gene pairs per cancer from the HEST-bench-trained precision matrix Σ_c^{-1} and cross-reference each against curated marker lists from PanglaoDB, CellMarker 2.0, and MSigDB Hallmark. Supplementary Table S2 lists the pairs that match a validated program, with the supporting reference; the *Type* column distinguishes cross-pathway matches (different gene families, same program), intra-family cell-type markers (e.g., keratins → keratinocyte), and generic housekeeping or cell-state axes that are reported but not claimed as cell-type structure. Pairs whose program assignment the external databases did not support are listed separately under *Excluded (unverified) assignments* below and excluded from the match counts.

Table S2: Top partial-correlation pairs matching validated marker programs, with supporting reference. Match counts per cancer: IDC 8/20 (40%), SKCM 6/20 (30%), LUNG 6/20 (30%), PRAD 2/20 (10%, housekeeping only).

Cancer	Gene pair	Program	Type	r
IDC	AQP1-VWF	endothelial	cross-pathway	+0.55
	KIT-TPSAB1	mast cell	cross-pathway	+0.50
	CD3E-TRAC	T-cell	intra (cell-type)	+0.44
	KRT14-KRT5	keratinocyte	intra (cell-type)	+0.43
	ADH1B-ADIPOQ	adipocyte	cross-pathway	+0.42
	ESR1-GATA3	ER/luminal	cross-pathway	+0.33
	KRT5-KRT6B	keratinocyte	intra (cell-type)	+0.33
	MYH11-MYLK	contractile	intra (module)	+0.29
SKCM	DSC1-KRT2	keratinocyte	cross-pathway	+0.36
	KRT16-KRT17	keratinocyte	intra (cell-type)	+0.34
	MGP-POSTN	fibroblast/CAF	cross-pathway	+0.31
	HAPLN1-MGP	fibroblast/CAF	cross-pathway	+0.27
	MMRN1-VWF	endothelial	cross-pathway	+0.23
	KRT16-KRT5	keratinocyte	intra (cell-type)	+0.22
LUNG	MARCO-MCEMP1	alveolar macrophage	cross-pathway	+0.52
	MKI67-TOP2A	cell cycle	intra (state)	+0.46
	COL5A2-CTSK	fibroblast/ECM	cross-pathway	+0.42
	CD79A-MZB1	B-cell	cross-pathway	+0.42
	CD163-MARCO	macrophage	cross-pathway	+0.29
	CDK1-TOP2A	cell cycle	intra (state)	+0.29
PRAD	RPL37A-RPS29	ribosomal	intra (housekeep.)	+0.77
	MT1E-MT1X	metallothionein	intra (housekeep.)	+0.50

Excluded (unverified) assignments. The following high-ranking pairs were previously labelled as program matches but are not supported by the external databases and are excluded from the counts above. For IDC, DST-MYLK (DST is a hemidesmosome/IF linker, not a smooth-muscle marker) and ANKRD30A/AQP3/CLIC6-MYBPC1 (the "milk-protein" cluster is invalid since MYBPC1 is a skeletal-muscle gene). For PRAD, SEC11C-SPON2 and SEC11C-XBP1 ("secretory ER" is not a standard program; SEC11C is a housekeeping signal-peptidase and SPON2 a secreted ECM/biomarker protein). For SKCM, MGP-VIM (VIM is a non-specific pan-mesenchymal marker) and LYZ-WDFY4 (WDFY4 is a cDC1 rather than a macrophage

gene). For LUNG, CAPN8-MALL and GPRC5A-MALL (MALL is not an airway club marker; CAPN8 is gastric, GPRC5A an alveolar AT1 marker) and CD163-PLA2G7 (PLA2G7 is absent from curated macrophage marker sets).

Interpretation. For IDC, SKCM, and LUNG the top pairs match validated known programs at 30-40%, or 25-57× above the random-pairing baseline (0.7%, 1.2%, and 0.9% respectively, 1000 iterations). The most conservative slice, namely cross-pathway pairs in which different gene families mark the same program and which gene-family similarity alone cannot recover, is 20% for each of the three. Same-family pairs that are genuine cell-type markers (keratins → keratinocyte; CD3E/TRAC → T-cell) are counted; generic housekeeping/cell-state axes (ribosomal, metallothionein, cell-cycle) and the contractile module are reported as intra-family and not claimed as cell-type structure. PRAD is the negative case: its only matches are housekeeping families, with no cell-type-specific program. We do not claim the prior discovers programs since it inherits them from the training expression data, but the well-above-chance overlap supports the claim that Σ_c^{-1} encodes biologically meaningful structure for these cohorts.

Supplementary Note 3 Extended hyperparameter robustness

The main text reports that a single dataset-level default is close to the per-baseline optimum for the two prior weights, and Figure 5 in the main text shows the $\lambda_{\text{bio}} \times \lambda_{\text{sp}}$ sensitivity heatmaps for all seven baselines. This Supplementary Note provides additional detail beyond the main-text figure, covering the extended per-baseline discussion of the three departures (SEPAL, STPath, STFlow) from the shared default and the additional robustness sweeps in k (Fig. S1) and η (Fig. S2, Table S3).

Extended $(\lambda_{\text{bio}}, \lambda_{\text{sp}})$ discussion. Four baselines, namely Ridge, ST-Net, GCN, and HisToGene, behave identically: Δ increases monotonically as either prior is strengthened and peaks in a single shared region at $(\lambda_{\text{bio}}, \lambda_{\text{sp}}) = (0.15, 0.2\text{-}0.3)$ on HEST-bench and $(0.10, 0.2\text{-}0.3)$ on STimage. The remaining three baselines, SEPAL, STPath, and STFlow, depart from this and all three departures follow a single principle: a baseline that already encodes a given structure strongly gains little from the matching prior and is actively hurt when that prior is pushed too far. SEPAL is a dedicated spatial-expression model whose graph already extracts most of the spatial signal, so the spatial prior is largely redundant. Its HEST optimum falls to $\lambda_{\text{sp}} \approx 0$, and on STimage cross-cohort the spatial smoothing over-corrects and drives Δ sharply negative at high λ_{sp}, with the optimum at $(0.05, 0.3)$. STPath relies on UNI foundation-model features that already encode cancer-conditional structure, so the biological prior over-corrects; it peaks at a markedly lower $\lambda_{\text{bio}} \approx 0.05$, and Δ turns negative for $\lambda_{\text{bio}} \geq 0.15$. STFlow is the strongest baseline overall and the only iterative sampler, with uniformly small gains on HEST and the same high-λ_{sp} over-projection as SEPAL on STimage. Being graph-based is not by itself sufficient to trigger this redundancy: GCN also operates on the spatial-neighbour graph, yet its lighter spatial modelling leaves residual structure for the prior to exploit, so it still benefits from $\lambda_{\text{sp}} = 0.3$ and is not an exception. The effect is therefore one of degree, determined by how strongly the architecture already captures the relevant structure, rather than one of model category.

The per-cohort default of $(\lambda_{\text{bio}}, \lambda_{\text{sp}}, k) = (0.15, 0.2, 8)$ for HEST and $(0.10, 0.2, 8)$ for STimage yields a Δ within 0.01 of the per-baseline best on 6 of the 7 HEST cells; STPath is the exception with a canonical $\Delta = +0.0002$ against a best $\Delta = +0.013$ (the aggregation artifact discussed in the main text). On STimage the default stays within 0.005 of the per-baseline best for Ridge, ST-Net, GCN, HisToGene, and STPath. SEPAL loses only 0.0014 mean PCC relative to its STimage best at $(0.05, 0.3)$, and STFlow's STimage optimum likewise sits at $(0.05, 0.3)$ with a small canonical-default gap consistent with its preference for lower λ_{bio} on the cross-cohort transfer.

Neighborhood-size k sensitivity. Supplementary Fig. S1 varies k at the canonical HEST-bench configuration $(\lambda_{\text{bio}}, \lambda_{\text{sp}}) = (0.15, 0.2)$ while keeping all other settings fixed. Across $k \in \{4, 6, 8, 12, 16, 20\}$, the multi-prior gain remains nearly flat for both ST-Net and HisToGene, with variation below 0.001. This indicates that the correction is not sharply tuned to the chosen neighbourhood size, and we retain $k = 8$ as the default because it lies in a stable near-optimal region and corresponds to a local Visium neighbourhood without introducing excessive

long-range smoothing.

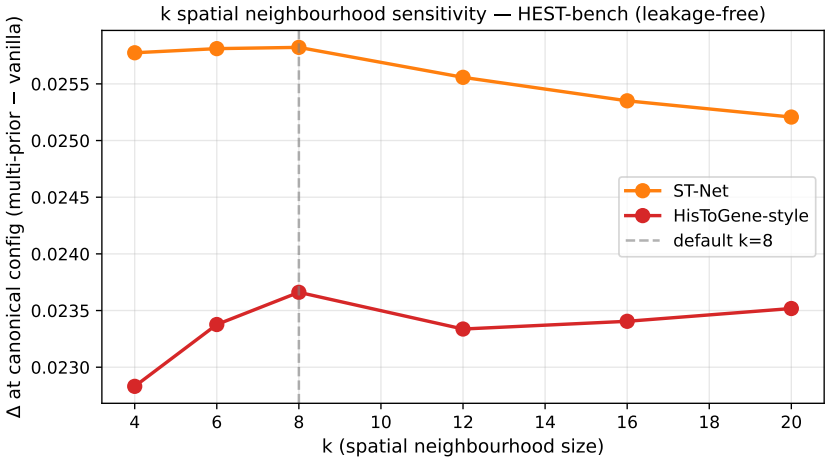

Figure S1: Sensitivity of the multi-prior correction to the spatial neighbourhood size k. Mean-PCC gain Δ is evaluated at the canonical HEST-bench configuration $(\lambda_{\text{bio}}, \lambda_{\text{sp}}) = (0.15, 0.2)$ while varying only k. The dashed line marks the default $k = 8$. Across $k \in \{4, 6, 8, 12, 16, 20\}$, Δ varies by less than 0.001 for both ST-Net and HisToGene baselines.

Covariance ridge η sensitivity. The biological prior corrects predictions with the per-cancer precision matrix $\Sigma_c^{-1} = (\Sigma_c + \eta I)^{-1}$, in which the ridge η is added only to keep the 50×50 empirical covariance invertible. To rule out dependence on this stabilization constant, we sweep $\eta \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ and measure the multi-prior gain Δ for every baseline while holding all other hyperparameters fixed (Supplementary Fig. S2, Supplementary Table S3). The per-η spread is at most 0.0032 (STFlow) and $\leq$ 0.0022 for the remaining six baselines, remaining below 1% of the absolute PCC change in every case. We fix $\eta = 10^{-3}$ a priori as the smallest value that keeps Σ_c stably invertible, rather than tuning it on the benchmark. For the five baselines with sizeable gains over 0.02 (Ridge, ST-Net, GCN, HisToGene, STPath), this spread is more than an order of magnitude smaller than the gain itself (11-27×). For the two baselines whose gains are intrinsically small (SEPAL, STFlow) the spread is proportionally larger relative to the gain ($\sim 2\times$), yet still tiny in absolute terms, and the gain remains strictly positive at every tested η.

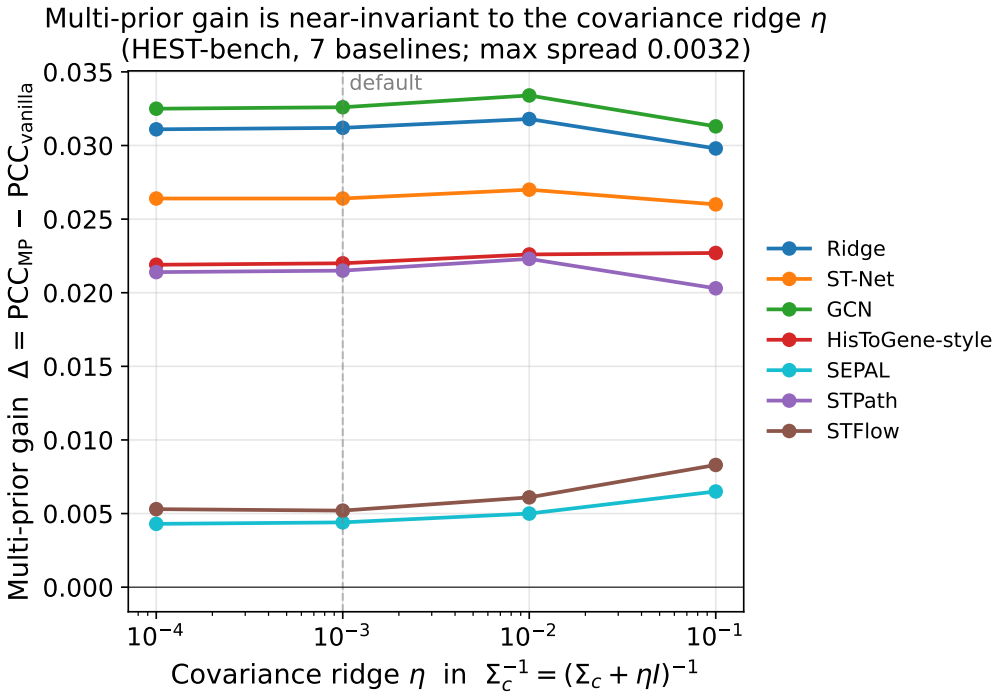

Figure S2: The multi-prior gain is near-invariant to the covariance ridge η. Multi-prior gain Δ versus the ridge η in $\Sigma_c^{-1} = (\Sigma_c + \eta I)^{-1}$ across all seven baselines (HEST-bench, $\lambda_{\text{bio}} = 0.15$, $\lambda_{\text{sp}} = 0.2$, $k_{\text{sp}} = 8$). Each line is one baseline (vanilla PCC in the legend); the dashed vertical line marks the default $\eta = 10^{-3}$. All curves are flat and remain strictly positive across the whole range: the largest per-η spread (max−min of Δ) is 0.0032 (STFlow), which is below 1% of the absolute PCC change.

Table S3: Robustness of the multi-prior gain Δ to the covariance ridge η. HEST-bench, $\lambda_{\text{bio}} = 0.15$, $\lambda_{\text{sp}} = 0.2$, $k_{\text{sp}} = 8$. STPath is evaluated per-(cancer, fold) here to make the η-sweep directly comparable across baselines; the headline Table 1 STPath value uses the slide-concatenated protocol reported in the original paper ($n = 10$ cancers, vanilla 0.3741). STFlow vanilla and Δ reported here use the 3-seed × 10-cancer × per-fold mean (87 raw per-PCC observations) for direct sensitivity comparison; the headline Table 1 STFlow value uses (cancer, fold)-averaged units after first averaging across seeds ($n = 29$ pairs).

Baseline	Vanilla PCC	Δ at ridge η				Spread
		10^{-4}	10^{-3}	10^{-2}	10^{-1}	(max−min)
Ridge	0.3179	+0.0311	+0.0312	+0.0318	+0.0298	0.0020
ST-Net	0.2792	+0.0264	+0.0264	+0.0270	+0.0260	0.0010
GCN	0.3473	+0.0325	+0.0326	+0.0334	+0.0313	0.0021
HisToGene	0.3620	+0.0219	+0.0220	+0.0226	+0.0227	0.0008
SEPAL	0.3611	+0.0043	+0.0044	+0.0050	+0.0065	0.0022
STPath	0.3754	+0.0214	+0.0215	+0.0223	+0.0203	0.0020
STFlow	0.3987	+0.0053	+0.0052	+0.0061	+0.0083	0.0032

Supplementary Note 4 STFlow temporal-schedule ablation

For iterative samplers such as STFlow, the multi-prior correction can be injected at each of the 5 denoising steps in the flow-matching sampler (sampling-time guidance) or applied only to the sampler output (post-hoc). Using the leakage-free prior estimated on HEST-bench, we ablate the timing of guidance injection by comparing several per-step schedules against post-hoc correction (Supplementary Fig. S3). The schedules differ only in the per-step weight trajectory; the correction form is identical across all of them.

Every per-step schedule outperforms post-hoc correction. The constant schedule yields a mean PCC of 0.4009, linear-increasing 0.4026, linear-decreasing 0.4014, late-only 0.4008, and cosine 0.4019, all exceeding both post-hoc variants (last-only and early-only, each 0.3988). The gap between the constant schedule and post-hoc correction is +0.0021 (paired Wilcoxon over the $n = 29$ cancer-fold cells, $p = 0.033$) and the strongest schedule, linear-increasing, exceeds post-hoc correction by +0.0038 ($p = 0.0007$). These results indicate that per-step guidance is beneficial for iterative samplers, whereas for non-iterative predictors post-hoc correction is the only applicable option and is functionally equivalent to the last step of a per-step schedule.

We adopt the parameter-free constant schedule as the headline configuration throughout. The two top schedules are statistically indistinguishable across most cells: linear-increasing exceeds constant by only +0.0017 in mean and +0.004 in median, winning in just 17 of 29 cells, and this slim margin is concentrated almost entirely in the single HEST READ cell (+0.0122), the one negative-Δ cell disclosed in the main text, while the remaining nine cancers are essentially tied. Selecting linear-increasing on the basis of this margin would therefore mean choosing, on the test folds, the schedule that happens to mask our sole disclosed negative result. The constant schedule avoids this since it is parameter-free, requires no test-fold selection, and follows the same fixed-default protocol applied to $(\lambda_{\text{bio}}, \lambda_{\text{sp}}, k)$.

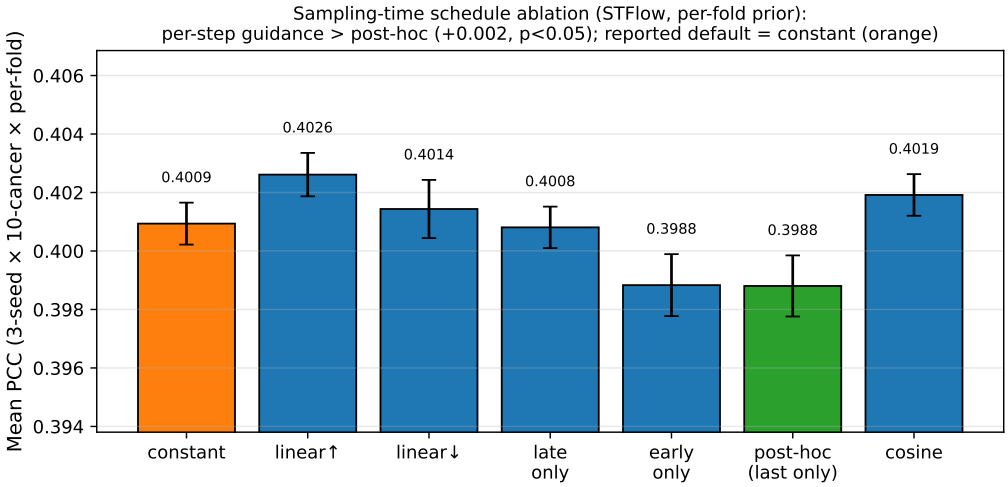

Figure S3: Sampling-time guidance schedules versus post-hoc correction applied to STFlow. Mean test PCC for STFlow's 5-step sampler under seven schedules for applying the multi-prior correction, over 3 seeds × 10 HEST-bench cancers × per-fold splits ($n = 29$ cancer-fold cells after seed-averaging for the paired Wilcoxon; the plotted means use the 3-seed × 10-cancer × per-fold aggregation for consistency with Supp Table S3); error bars are ±1 s.d. across seeds. Schedules differ only in which denoising steps receive the correction; the prior ($\lambda_{\text{bio}} = 0.15$, $\lambda_{\text{sp}} = 0.2$, $k = 8$) is identical across all bars. All per-step schedules exceed post-hoc (last-only); the constant-versus-post-hoc gap is +0.0021 (paired Wilcoxon, $p = 0.033$).

Supplementary Note 5 STimage-bench construction

STimage-1K4M is a public collection of $\sim$ 1,478 spatial transcriptomics slides re-encoded into the HEST-bench h5 schema. We construct STimage-bench, a per-cancer benchmark that mirrors HEST-bench's protocol, by filtering STimage-1K4M according to: (i) species is human; (ii) the "involves_cancer" flag is true; (iii) technology is Visium or VisiumHD, which excludes older ST platforms with too few genes; (iv) total gene count ≥ 2000; (v) coverage of the corresponding HEST-bench cancer's 50 HVG is $\geq 25/50$; (vi) sample ID is not in HEST-bench's 72-slide set (no leakage). For each retained sample, the slide is grouped into the HEST-bench cancer whose tissue criterion it matches, with "breast" tissue mapping to IDC or LYMPH_IDC and "kidney" tissue mapping to CCRCC. The resulting STimage-bench has per-cancer slide counts shown in Supplementary Table S4. LUNG and PRAD are absent from STimage-bench because no STimage-1K4M slides match our tissue and HVG-coverage criteria. We perform 3-fold cross-validation per cancer (24 folds total).

Table S4: STimage-bench per-cancer slide counts. Slides retained after filtering, grouped into the matching HEST-bench cancer. The IDC count is capped at 60 (from 205) to limit cohort imbalance. LUNG and PRAD are absent.

Cancer	Slides (cap 60)
CCRCC	24
COAD	4
HCC	12
IDC	60 (cap 205)
LYMPH_IDC	60
PAAD	18
READ	4
SKCM	4
Total	186

Tolerant adata loading. STimage-bench slides with partial HVG coverage are loaded via a tolerant loader that filters the requested gene list to those present in the slide and zero-pads the rest. This is necessary for the spatial-transcriptomics convention of treating any missing gene as expression 0. A consequence is that STimage-specific bio priors $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ become contaminated by zero-padded columns, which is why we use HEST-bench-trained priors for all cross-cohort evaluation (Methods).

Supplementary Note 6 Statistical test details

To evaluate the performance improvement Δ, we report paired Wilcoxon signed-rank tests with Bonferroni correction and 95% bootstrap confidence intervals on the mean improvement. We record below the exact sampling unit and effective n for each (baseline, cohort) pair.

Sampling unit per cell. For Ridge, ST-Net, GCN, HisToGene, and SEPAL the per-(cancer, fold) PCC is computed from the per-fold prediction. HEST-bench has 29 folds across 10 cancers and STimage has 24 folds across 8 cancers. For STFlow on HEST-bench, the PCC is computed for each (cancer, fold, seed) triple, giving 87 raw per-PCC observations across 3 seeds, 10 cancers, and the 29 (cancer, fold) splits of the HEST-bench cross-validation. For statistical testing, the per-seed PCC values are first averaged within each (cancer, fold) pair, and the (cancer, fold) pair is then treated as the paired unit, yielding $n = 29$ paired observations for the Wilcoxon signed-rank test and the bootstrap CI. For STFlow on STimage the score is computed per-(cancer, fold), with $n = 16$ paired observations (6 cancers $\times$ 3 folds, minus 2 COAD folds lacking test slides), since READ and SKCM are excluded because all their slides have partial HVG coverage. For STPath the PCC is computed once per cancer under the original slide-concatenated protocol, with $n = 10$ on HEST and $n = 8$ on STimage.

Wilcoxon test. Paired Wilcoxon signed-rank. The raw p-value is Bonferroni-corrected by multiplying by 7 (the number of baselines per cohort) and clipping to $[0, 1]$.

Bootstrap CI. 1000 iterations resampling the paired observations with replacement. Reported CI is the 2.5-97.5 percentile of the bootstrap mean-Δ distribution at seed 42.