Beyond Counts: A Distributional Robustness Margin For Pathology Foundation Models

Clément Grisi*[1], Jeroen van der Laak[1], and Geert Litjens[1]

[1]Department of Pathology, Radboud University Medical Center, Nijmegen, The Netherlands

Abstract

Pathology foundation models are approaching clinical deployment, yet remain vulnerable to systematic non-biological variation across centres. Differences in tissue preparation, staining and scanning are strongly encoded in their representations, enabling shortcut learning and weakening generalisation across cohorts and institutions. The Robustness Index (`RI`) quantifies whether local representation geometry is dominated by biology or by non-biological variation, but its count-based formulation discards distance information. We show that adding distance weights changes little because the deeper limitation lies in `RI`'s pooled, fixed-neighbourhood design, which obscures sample-level heterogeneity and effectively evaluates only a model-dependent subset of samples. We introduce the Cross-confounder Robustness Margin (`CRoMa`), a sample-resolved measure that directly compares distances to cross-confounder biological matches and same-confounder biological distractors. By design, `CRoMa` recasts robustness as a cohort-wide margin distribution rather than a single pooled score. We evaluated frozen representations from 20 tile-level encoders across three benchmarks and 4 slide-level encoders on a fourth. Rankings by median `CRoMa` were broadly consistent across datasets, while the underlying distributions revealed substantial within-model heterogeneity. Every tile encoder retained a confounder-dominated lower tail, whose prevalence and severity varied markedly across models. These distinct robustness profiles frame model selection as a Pareto trade-off between typical and lower-tail robustness. Higher `CRoMa` margins were also associated with smaller shortcut-induced performance drops after supervised adaptation. By turning representation geometry into a distributional robustness readout that anticipates downstream shortcut susceptibility, `CRoMa` provides a principled basis for robustness assessment and model selection.

1 Introduction

Advances in self-supervised learning have ushered in the era of foundation models. By pretraining on large collections of unlabeled data, these models shift much of the learning burden to the representation-learning stage, producing general-purpose embeddings that can be adapted to downstream tasks with substantially fewer labels [15, 4, 1, 27]. This is especially compelling in medical domains, where expert annotation is costly and disease prevalence is often long-tailed, leaving many clinically important tasks without sufficient data for proper supervised training. In computational pathology, this promise is beginning to materialize: models built on foundation-model representations have started to match or surpass the narrower supervised systems that preceded them [5, 33, 31, 3, 26].

*corresponding author: clement.grisi@radboudumc.nl

The rapid development of pathology foundation models has, however, outpaced our ability to assess their limitations rigorously [23]. A central vulnerability is their sensitivity to batch effects introduced by tissue preparation, sectioning, staining, and whole-slide scanning, which are strongly encoded in their feature spaces [20, 6, 9, 19]. This vulnerability follows from how these models are pretrained: self-supervised learning rewards representations that capture recurring structure in the data, allowing technical artifacts that alter the appearance of pathology images to be encoded alongside biology. Such co-encoding becomes problematic when technical variation aligns with clinical labels. If, for example, one institution contributes mostly malignant cases and another mostly benign cases, a downstream model can exploit institution-specific staining or scanner signatures as shortcuts. In that setting, predictions may reflect acquisition artifacts rather than morphology, undermining generalisability across cohorts and institutions [12, 14, 7]. Robustness to these confounders is therefore a core requirement for developing models that can support reliable clinical deployment.

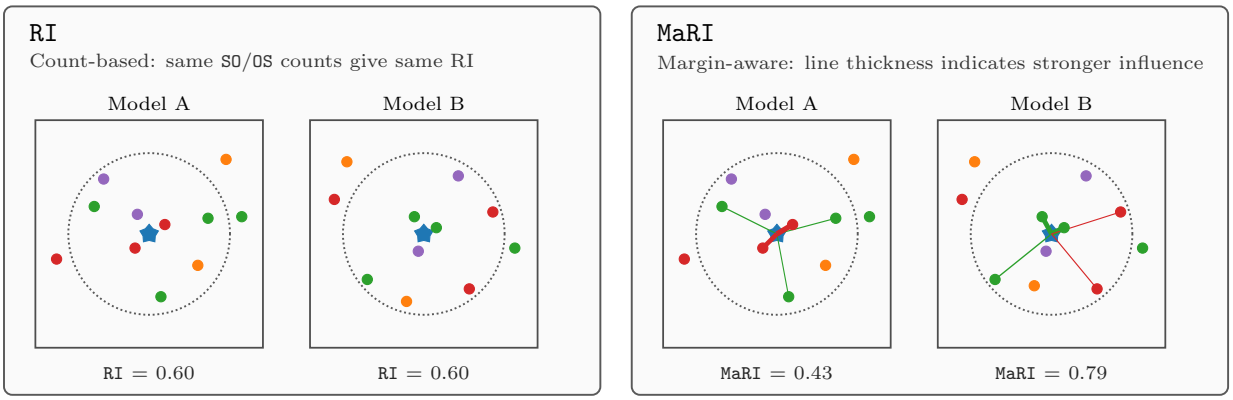
PathoROB recently formalized this problem by introducing a public benchmark for assessing pathology foundation model robustness to non-biological variation [19]. Its central metric, the Robustness Index (`RI`), asks whether local neighborhoods in representation space are organized primarily by biology or by confounder. For each anchor, that is, each sample whose local neighbourhood is being scored, `RI` compares cross-confounder biological matches with same-confounder biological distractors, yielding an interpretable count-based readout of local organization. However, this design has three structural limitations. First, because `RI` is count-based, it discards geometry: two models can have identical neighbour counts while placing biological matches and confounder-driven distractors at very different distances from the anchor. Second, because `RI` is ultimately reported as a pooled aggregate, it compresses heterogeneous sample-level behaviour into a single score, obscuring vulnerable tails, localized failure modes, and confounder-specific fragility. Third, `RI` is defined only for anchors whose fixed-k neighbourhood contains the typed neighbours needed to form the biological-versus-confounder contrast. Anchors lacking such neighbours are silently excluded from the pooled score: different models can be evaluated on different effective subsets of the data, complicating cross-model comparisons.

To address these limitations, we introduce the Cross-confounder Robustness Margin (`CRoMa`), a geometry-aware measure of model robustness in representation space. Rather than counting neighbors within a fixed neighborhood, `CRoMa` measures, for each sample, the distance margin between cross-confounder biological matches and same-confounder distractors. This formulation requires no per-model neighborhood tuning and is defined on every sample, enabling finer-grained analyses that can reveal confounder-dominated subgroups obscured by pooled averages. We evaluate `CRoMa` across 20 tile-level and 4 slide-level pathology foundation models on four benchmarks. Model rankings are broadly consistent across datasets ($\rho \approx 0.90$), and `CRoMa` closely tracks downstream shortcut learning, supporting its relevance for deployment-oriented assessment of foundation model robustness.

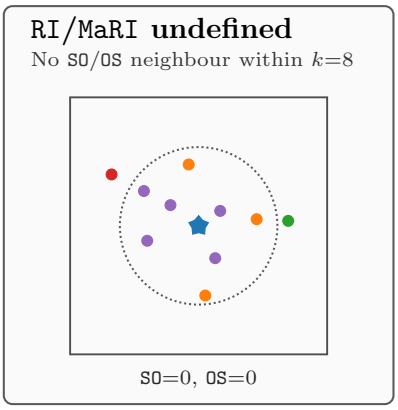
2 Methods

We compare three representation-level robustness metrics: the Robustness Index (`RI`), its distance-weighted refinement (`MaRI`), and the Cross-confounder Robustness Margin (`CRoMa`). All three metrics are built from the same typed-neighbour evidence. For each anchor sample, they compare

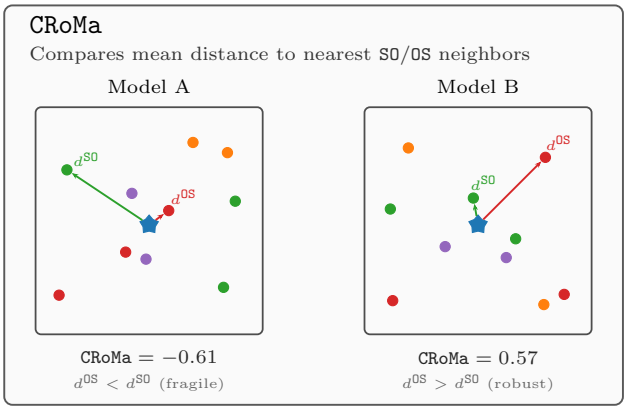
cross-confounder biological matches, which share the anchor's biological label but differ in confounder, with same-confounder biological distractors, which share the anchor's confounder but differ in biological label. We first define this common setup, then show how each metric aggregates the resulting evidence: `RI` as a pooled count of typed neighbours, `MaRI` as a pooled distance-weighted analogue, and `CRoMa` as a per-sample signed margin between the two neighbour types. Figure 1 illustrates the three metrics on a single worked neighbourhood.



(a) Count-identical neighbourhoods can differ in margin.



(b) Fixed-k metrics can be undefined.



(c) `CRoMa` compares nearest typed distances.

Figure 1: **Count- and distance-based robustness metrics. SO** neighbours share biology but differ in confounder, whereas **OS** neighbours differ in biology but share the confounder. **a)** `RI` counts typed neighbours within a fixed-k neighbourhood and gives both models the same score. `MaRI` weights the same evidence by distance (edge width denotes weight), resuling in different scores for each model. **b)** If the fixed neighbourhood contains neither typed neighbors, `RI` and `MaRI` are undefined for that sample, which does not contribute to the pooled score. **c)** `CRoMa` compares the mean distance to the nearest typed neighbours, yielding a value for every sample by construction. Positive values denote biology-dominant geometry and negative values confounder-dominant geometry.

2.1 Setup and notation

Each sample i is represented by a feature vector z_i extracted using a foundation model, a biological label y_i, a confounder label c_i (e.g. its acquisition centre), and a grouping identifier g_i denoting the physical unit from which the sample was derived (a whole-slide image for tile-level samples, a case for slide-level samples). We refer to the sample being scored as the *anchor*. Features are L2-normalized, and candidate neighbours are ranked by cosine distance $d_{ij} = 1 - \cos(z_i, z_j)$. Before scoring, samples from the same physical unit as the anchor ($g_j = g_i$) are excluded from its candidate neighbourhood, preventing any metric from benefiting from near-duplicate samples, such as tiles from the same slide.

For an anchor i and neighbour j, the pair is *typed* according to whether biology and confounder match, following the four-way scheme introduced by [6] and adopted in PathoROB [19]. Two pair types provide the relevant contrast. **SO** pairs have the **S**ame biological label and an **O**ther confounder ($y_j = y_i, c_j \neq c_i$). They are cross-confounder biological matches and therefore support a biology-organized representation when close to the anchor. **OS** pairs have **O**ther biology and the **S**ame confounder ($y_j \neq y_i, c_j = c_i$). They are same-confounder distractors and support a confounder-organized representation when close. The remaining pair types are uninformative and are therefore excluded from scoring. For both **SS** and **OO** neighbors, their proximity to the anchor cannot be attributed specifically to either biological or confounder signal: **SS** pairs share both factors, whereas **OO** pairs share neither. Accordingly, all three metrics compare only **SO** and **OS** neighbours.

2.2 Robustness Index

We follow the PathoROB definition of the Robustness Index (`RI`) [19]. For a fixed neighbourhood size k, let $\mathcal{N}_k(i)$ denote the k nearest neighbours of anchor i, after excluding samples from the same physical unit. Within this neighbourhood, `RI` counts the two informative neighbour types:

$$\texttt{SO}_i = \sum_{j \in \mathcal{N}_k(i)} \mathbf{1}[y_j = y_i \wedge c_j \neq c_i] \tag{1}$$

$$\texttt{OS}_i = \sum_{j \in \mathcal{N}_k(i)} \mathbf{1}[y_j \neq y_i \wedge c_j = c_i] \tag{2}$$

A model is then summarized by the pooled proportion of informative neighbours that are cross-confounder biological matches:

$$\texttt{RI} = \frac{\sum_i \texttt{SO}_i}{\sum_i (\texttt{SO}_i + \texttt{OS}_i)} \tag{3}$$

The score ranges from 0, indicating neighbourhoods dominated by same-confounder distractors, to 1, indicating neighbourhoods dominated by cross-confounder biological matches. Unless stated otherwise, and to stay consistent with the original PathoROB protocol, we adopt its two-step procedure for selecting k. First, for each model, we chose the value of k that maximizes biological k-nearest-neighbour balanced accuracy. We then evaluate all models using the median of these model-specific optimal values. This choice supports a common evaluation scale, but is not free of ambiguity: there is no uniquely principled choice of k that is both model-specific and directly comparable across models. The l slide-level panel is an explicit exception and is evaluated at per-model $k^\star$, as detailed below and in Supplementary Section A.12.

2.3 Margin-aware Robustness Index

Because RI is count-based, it treats all neighbours within the fixed-k set equally, regardless of their distance from the anchor. Two models can therefore obtain identical RI scores while inducing opposite local geometries: in one, cross-confounder biological matches lie much closer to the anchor than same-confounder distractors, whereas in the other, the distractors lie closer than the biological matches (Figure 1a). Because RI considers only neighbor counts, it cannot distinguish between these two models. Only a distance-weighted measure can. The margin-aware robustness index (MaRI) provides a minimal distance-weighted refinement: it preserves the same fixed-k, proportion-based formulation, but weights each typed neighbour by its distance to the anchor:

$$w_{ij} = \exp\left(-\frac{d_{ij}}{\tau}\right) \tag{4}$$

where $\tau > 0$ is a temperature parameter. The weighted typed evidence becomes:

$$\texttt{SO}_w(i) = \sum_{j \in \mathcal{N}_k(i)} w_{ij}\, \mathbf{1}[y_j = y_i \wedge c_j \neq c_i] \tag{5}$$

$$\texttt{OS}_w(i) = \sum_{j \in \mathcal{N}_k(i)} w_{ij}\, \mathbf{1}[y_j \neq y_i \wedge c_j = c_i] \tag{6}$$

resulting in the pooled index:

$$\texttt{MaRI} = \frac{\sum_i \texttt{SO}_w(i)}{\sum_i \left(\texttt{SO}_w(i) + \texttt{OS}_w(i)\right)} \tag{7}$$

Smaller values of τ concentrate the score on the closest typed neighbours. As τ grows the weights flatten and MaRI approaches RI's unweighted counts. To keep the temperature on the scale of the distances being weighted, we set τ separately for each model to the median distance among typed **SO**/**OS** neighbours.

Distance weighting changes how typed neighbours contribute to the score, but it does not resolve the absence of typed evidence within the pre-specified neighbourhood. MaRI therefore inherits a structural limitation of RI: an anchor i contributes to the pooled score only if its fixed-k neighbourhood contains at least one informative typed neighbour, such that $\texttt{SO}_i + \texttt{OS}_i > 0$ (Figure 1b). Anchors that do not satisfy this condition are not counted as failures but are silently excluded from the aggregate. Consequently, pooled RI and MaRI are computed over a model-dependent subset of the evaluation cohort: the anchors for which at least one typed neighbour falls within the selected neighbourhood. Because this subset can vary across models, nominally comparable scores may in fact summarize different groups of anchors, undermining direct cross-model comparison. This limitation is structural rather than incidental, arising directly from the criterion used to select k (Supplementary Section A.4).

2.4 Cross-confounder Robustness Margin (CRoMa)

CRoMa keeps the distance sensitivity introduced by MaRI while ensuring that every samples gets a score. Rather than requiring informative neighbours to appear within a pre-specified k-neighbourhood, it locates the nearest informative neighbours of each type and measures their distance separation directly (Figure 1c). For each sample i, let $d_m^{\texttt{SO}}(i)$ and $d_m^{\texttt{OS}}(i)$ denote the mean cosine distances to its m nearest **SO** and m nearest **OS** neighbours:

$$d_m^{\texttt{SO}}(i) = \frac{1}{m}\sum_{j=1}^{m} d_{i,\,\sigma_j^{\texttt{SO}}(i)} \tag{8}$$

$$d_m^{\texttt{OS}}(i) = \frac{1}{m}\sum_{j=1}^{m} d_{i,\,\sigma_j^{\texttt{OS}}(i)} \tag{9}$$

where $\sigma_j^{\texttt{SO}}(i)$ and $\sigma_j^{\texttt{OS}}(i)$ index the j-th nearest **SO** and **OS** neighbours, respectively.

The per-sample robustness margin is defined as:

$$\texttt{CRoMa}_m(i) = \frac{d_m^{\texttt{OS}}(i) - d_m^{\texttt{SO}}(i)}{d_m^{\texttt{OS}}(i) + d_m^{\texttt{SO}}(i)} \tag{10}$$

This yields a signed, scale-free margin in $(-1, 1)$. Positive values indicate that same-confounder biological distractors are farther from the anchor than cross-confounder biological matches, consistent with a biology-dominant local geometry. Negative values indicate a confounder-dominant geometry. Averaging over m neighbours per type reduces sensitivity to single-neighbour outliers. In contrast to the neighbourhood size k used by `RI` and `MaRI`, which can vary between models and benchmarks, m is fixed a priori and used unchanged across all models and datasets. We set $m = 5$, ensuring that no individual neighbour contributes more than one fifth of a typed mean. Our conclusions are stable across a sweep of m values (Supplementary Section A.3).

Because `CRoMa` depends on the distance ratio $r_i = d_m^{\texttt{OS}}(i)/d_m^{\texttt{SO}}(i)$ rather than on absolute distances[1], it is invariant to any global multiplicative rescaling of distances within a model. Thus, a model is not favoured simply because its embedding distances are uniformly larger or smaller: only the relative separation between cross-confounder biological matches and same-confounder biological distractors matters. This gives `CRoMa` a common interpretation across models and datasets, while retaining a direct geometric meaning. For example, a median `CRoMa` of 0.7 implies that, for the typical anchor, the nearest same-confounder biological distractors are approximately 5.7× farther away than the nearest cross-confounder biological matches[2]. A complete geometric interpretation is provided in Supplementary Section A.5.

By construction, `CRoMa` assigns a score to every sample, provided that at least m `SO` and m `OS` neighbours exist in the evaluation set, a condition satisfied by all benchmarks used in this study for $m = 5$. The resulting margins $\texttt{CRoMa}_m(i)$ therefore define an empirical robustness distribution over the full evaluation cohort, rather than over the model-dependent support set on which `RI` and `MaRI` are effectively computed. This distribution is the primary robustness readout, because it preserves sample-level heterogeneity that a pooled score necessarily compresses. For compact model comparison, we summarize it along three complementary axes: central tendency, the prevalence of confounder-dominant samples, and the severity of robustness failures.

As a robust measure of central tendency, we report the median robustness margin:

$$\texttt{CRoMa} = \text{median}_i\, \texttt{CRoMa}_m(i) \tag{11}$$

[1] $\texttt{CRoMa}_m(i) = (r_i - 1)/(r_i + 1)$

[2] At $\texttt{CRoMa}_m(i) = 0.7$, $r_i = (1 + \texttt{CRoMa}_m(i))/(1 - \texttt{CRoMa}_m(i)) = 1.7/0.3 \approx 5.7$

The sign of the median indicates whether the typical sample occupies a biology-dominant or confounder-dominant local geometry. Yet the median alone obscures heterogeneity: two models may have similar median margins but differ markedly in the prevalence and magnitude of adverse samples. Because deployment risk is often driven by such samples, we additionally characterise the lower end of the CRoMa distribution along two axes: *prevalence* and *severity*.

Prevalence is measured at the decision boundary. Let F denote the empirical cumulative distribution function of $\texttt{CRoMa}_m(i)$. We report:

$$F(0) = \mathbb{P}\left[\texttt{CRoMa}_m(i) < 0\right], \tag{12}$$

Since zero is the meaningful boundary between biology-dominant and confounder-dominant geometry, $F(0)$ provides the fraction of samples whose local neighbourhood is dominated by the confounder. Severity is measured in the lower tail of the margin distribution. Let Q_α denote the α-quantile of $\texttt{CRoMa}_m(i)$. We define the lower-tail mean as:

$$\texttt{LTM}_\alpha = \mathbb{E}[\texttt{CRoMa}_m(i) \mid \texttt{CRoMa}_m(i) \leq Q_\alpha], \tag{13}$$

$\texttt{LTM}_\alpha$ measures the average robustness margin among the worst α fraction of samples. We report $\texttt{LTM}_\alpha$ at $\alpha = 0.10$, corresponding to the mean of the worst decile. Our conclusions are stable across $\alpha \in 0.05, 0.10, 0.20$ (Supplementary Section A.2).

2.5 Benchmarks

We evaluate all metrics on benchmarks that pair a biological label with the acquisition site that produced the sample (Table 1; per-cell cardinalities in Figure 2). Three tile-level benchmarks are taken directly from PathoROB [19]: Camelyon, comprising breast lymph-node tumour and normal tissue across two centres, TCGA, comprising cancer types sampled across centres, and Tolkach-ESCA, comprising six oesophageal tissue classes across three cohorts. We refer to the original PathoROB study for detailed dataset descriptions. To assess whether the same biology–confounder tension extends beyond tile-level encoders, we additionally curate PcaBiop, a slide-level benchmark of H&E prostate biopsies sourced from PANDA [2]. Together, these benchmarks span multiple representation scales, organs, biological label structures, and confounding regimes. Representative samples from each benchmark are shown in Supplementary Figure 7, illustrating the visually apparent non-biological variation targeted by these metrics.

Benchmark	Level	Biological label (#cls)	Confounder (#)	Samples
Camelyon [19]	tile	breast LN: tumour/normal (2)	medical centre (2)	20,400 tiles
TCGA (4×4) [19]	tile	cancer type (4)	medical centre (4)	5,760 tiles
Tolkach-ESCA [19]	tile	oesophageal tissue (6)	medical centre (3)	9,000 tiles
PCaBiop	slide	prostate: benign/cancer (2)	medical centre (2)	1,000 slides

Table 1: **Evaluation benchmarks span multiple organs and representation scales.** The four cohorts cover multiple organs and biological conditions, as well as tile- and slide-level representations. Confounders are non-biological acquisition artifacts from the medical centre that contributed the sample. Camelyon, TCGA-4x4 and Tolkach-ESCA are taken from PathoROB [19]. PCaBiop is a collection of prostate biopsises sourced from PANDA [2].

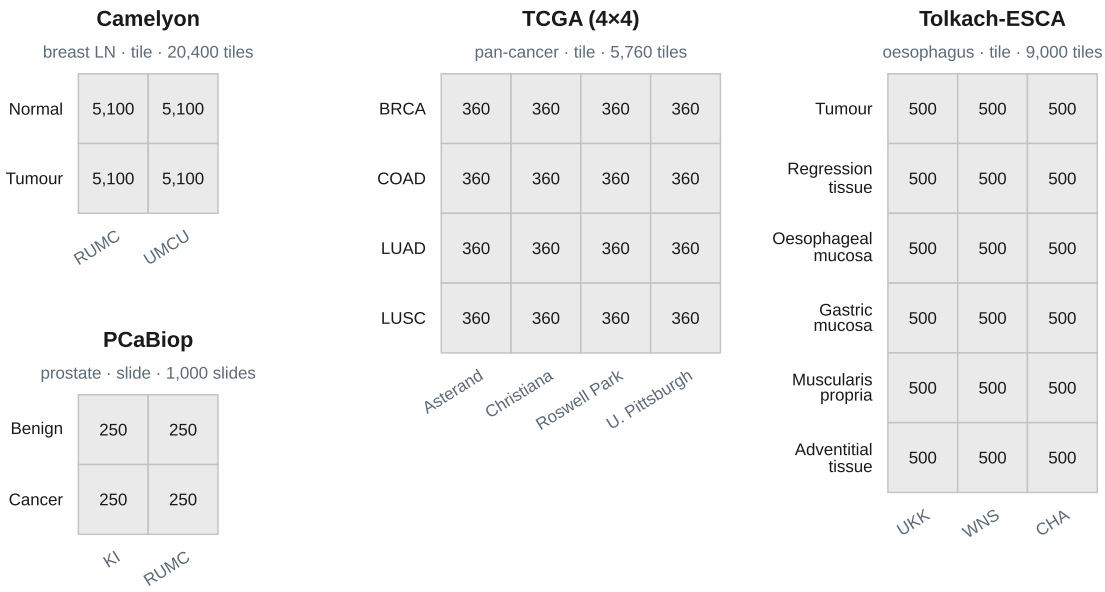


Figure 2: **Each benchmark is balanced across biological class and confounder.** Cells show evaluated samples per biological-class×confounder combination: 5,100 tiles for Camelyon (20,400 total), 360 for TCGA-4x4 (5,760), 500 for Tolkach-ESCA (9,000) and 250 slides for PCaBiop (1,000). Rows denote biological classes and columns acquisition centres. The paired TCGA-2 × 2 design is shown in Supplementary Figure 16.

2.6 Models

All models are evaluated as frozen feature extractors, so the metrics probe the representations exposed after pretraining rather than any task-specific adaptation (Table 2). For the tile-level benchmarks, we evaluate a common panel of 20 tile encoders spanning a broad range of architecture sizes, pretraining corpus scales, and self-supervised objectives, alongside a natural-image control (`DINOv2-B`) reported separately throughout. For the slide-level benchmark, we evaluate 4 whole-slide encoders that aggregate tile-level information into a single slide embedding. Because robustness estimates may be affected by overlap between a benchmark and a model's pretraining domain, we record each model's disclosed pretraining corpus, to the best of our ability, in Table 2.

2.7 Evaluation protocol

The evaluation protocol is fixed across models within each benchmark, with one exception to the shared operating-point rule. On the tile-level benchmarks, `RI` and `MaRI` use the median-k protocol: each model's biological $k^\star$ is identified first, after which all models are evaluated at the median of these optima. Since the slide-level panel contains only four encoders, the shared lower median (the lower of the two central optima) collapses to $k = 3$ and reduces mean support from 36.9% to 27.0%. We therefore evaluate slide-level models at each encoder's biologically selected $k^\star$. This exception affects `RI`, `MaRI`, their support and biological k-NN accuracy, whereas `CRoMa` and its distributional summaries remain k-independent. `MaRI` additionally uses a per-model temperature τ, and `CRoMa` uses $m = 5$ neighbours per type for all models and benchmarks. For TCGA, PathoROB provides two native evaluation configurations. We focus on the TCGA-4x4 configuration – four cancer

Model	Method	Params	WSIs / tiles	Dim	Pretraining corpus
Tile-level encoders					
Virchow2 [35]	DINOv2	632M	3.1M / 1.9B	2560	MSKCC (prop.)
Virchow [31]	DINOv2	632M	1.5M / 2B	2560	MSKCC (prop.)
UNI2-h [24]	DINOv2	681M	>350k / >200M	1536	MGB (prop.)
UNI [5]	DINOv2	307M	100k / 100M	1024	MGB (prop.)
CONCHv1.5 [8]	vision–language	307M	100k / 100M	768	PMC-OA + educational
CONCH [21]	vision–language	86M	21.4k / 16M	512	PMC-OA + educational
H-optimus-1 [29]	n/d	1.1B	>1M / n/d	1536	Bioptimus (prop.)
H-optimus-0 [28]	DINOv2	1.1B	500K / 273M	1536	Bioptimus (prop.)
H0-mini [9]	distilled (H-opt-0)	86M	6,093 / 43M	1536	**TCGA (distilled)**
Prov-GigaPath [33]	DINOv2	1.1B	171k / 1.3B	1536	Providence (prop.)
Midnight-12k [17]	DINOv2 (var.)	1.1B	12k / 384M	3072	**TCGA (public)**
Prost40M [13]	DINO	22M	2k / ∼40M	384	**TCGA-PRAD + LEOPARD**
Phikon [10]	iBOT	86M	6k / 43.3M	768	**TCGA (public)**
Phikon-v2 [11]	DINOv2	307M	∼58k / ∼456M	1024	**PanCancer-XL (incl. TCGA)**
Hibou-L [25]	DINOv2	307M	>1M / 1.2B	1024	HistAI (prop.)
Hibou-B [25]	DINOv2	86M	>1M / 512M	768	HistAI (prop.)
mSTAR [34]	distilled (UNI)	307M	11,765 / >116M	1024	**TCGA (self-taught)**
GPFM [22]	distilled (multi-expert)	307M	72,280 / 190M	1024	**33 public (TCGA, CAMELYON)**
MUSK [32]	vision–language	307M	∼33k / 50M	2048	**TCGA + PMC-OA + Quilt-1M**
GenBio-PathFM [16]	JEDI (DINO+JEPA)	1.13B	∼177k / ∼400M	4608	**HistAI + TCGA + GTEx + REG**
DINOv2-B [27]	DINOv2	86M	n/a	768	LVD-142M (natural images)
Slide-level encoders					
PRISM [30]	Perceiver	99M*	587k / n/d	1280	Paige (prop.)
TITAN [8]	CoCa	48.5M	336k / n/d	768	MGB (prop.)
Prov-GigaPath [33]	LongNet	86.3M*	171k / 1.3B	768	Providence (prop.)
MOOZY [18]	grid ViT	85.8M‡	77k‡ / 1.7B	768	**56 public (TCGA, PANDA)**

Table 2: **Foundation models overview.** 'n/d' undisclosed and 'n/a' not applicable. Boldface pretraining corpora include TCGA. * Slide-encoder parameter count as reported by Ding et al. [8]. `CONCHv1.5` lists the corpus of its vision trunk, which is initialised from `UNI`; its vision–language stage additionally used 1.26M image–caption pairs. `Midnight-12k` draws tiles online during training, so 384M counts tile draws rather than unique tiles. ‡ `MOOZY` is pretrained on slide feature grids rather than on slides directly: a slide available at two magnifications contributes one grid per magnification (53,286 at 20× and 23,848 at 40×), so the number of distinct WSIs is not separately reported. Its parameter count comprises a 64.1M slide-and-case encoder and a frozen 21.7M ViT-S patch encoder.

types (BRCA, COAD, LUAD, LUSC) across four medical centres. For completeness, we additionally report results on PathoROB's paired TCGA-2x2 benchmark (Supplementary Section A.11). Alongside `RI` and `MaRI`, we report each model's *support*, defined as the fraction of samples for which $\texttt{SO}_i + \texttt{OS}_i > 0$ within the pre-specified k neighbourhood. `CRoMa` is defined on the full evaluation cohort by construction. Metric implementations and evaluation manifests are released as an open-source Python package, `croma`.

3 Results

We use Camelyon as the primary benchmark and then examine the consistency of the findings across datasets and representation scales. Camelyon provides the clearest stress test for the proposed metrics. It is the largest tile-level cohort considered here (20,400 tiles, compared with 9,000 for Tolkach-ESCA and 5,760 for TCGA; Table 1), giving the most stable basis for per-sample margins and lower-tail analyses. It also spans the broadest robustness range: encoders extend from clearly

biology-dominant to clearly confounder-dominant, whereas the other tile-level benchmarks place all models in the biology-dominant regime (Section 3.4). Finally, it exposes most sharply the central limitation of fixed-k scores, with support reaching its lowest range ($10 - 46\%$, compared with $67 - 100\%$ on Tolkach-ESCA and $99 - 100\%$ on TCGA-4x4).

3.1 Distance weighting exposes the limits of fixed-k robustness scores

The Robustness Index (`RI`) evaluates robustness from the composition of a fixed k-nearest-neighbour set. For each sample, it counts neighbours with the same biological label but a different confounder (**SO**) against neighbours with a different biological label but the same confounder (**OS**; Methods). This construction captures *how many* favourable and unfavourable neighbours appear within the selected neighbourhood, but it is insensitive to *where* they appear. A cross-confounder biological match immediately adjacent to the anchor and one lying near the edge of the neighbourhood contribute identically. The same holds for same-confounder distractors.

`MaRI` was designed as the minimal geometric correction to this count-based score. It preserves the fixed-k evaluation protocol and the proportional **SO**/(**SO**+**OS**) structure of `RI`, but weights typed neighbours by their distance to the anchor (Methods). Thus, samples with identical typed-neighbour counts can receive different scores when favourable and unfavourable neighbours occupy different positions in the local geometry. Empirically, this correction is coherent but modest (Table 3). On Camelyon, Δ = `MaRI` − `RI` remains within ± 0.07 for every encoder, and model rankings are nearly unchanged (Spearman $\rho = 0.99$; Table 9). The sign of Δ, however, is informative: it is positive when favourable `SO` neighbours are systematically closer than same-confounder distractors (`UNI2-h`, $+0.033$; `Virchow2`, $+0.017$), and negative when distractors are closer (`Midnight-12k`, -0.070; `Virchow`, -0.043).

Yet the close agreement between pooled `RI` and `MaRI` should not be interpreted as evidence that distance carries little information. It is instead a consequence of the fixed-k, pooled design, which imposes two separable constraints. First, fixed neighbourhoods act as a support filter: a sample contributes to the pooled score only if its selected neighbourhood contains at least one **SO** or **OS** neighbour. Otherwise, it provides no **SO**/**OS** contrast and is silently excluded, making the effective evaluation cohort model-dependent, which complicates cross-model comparisons. Second, by pooling over the samples that remain, `MaRI` departs from `RI` only when **SO** neighbours are, *in aggregate*, systematically closer (or farther) than **OS** neighbours. Per-sample asymmetries of opposite sign cancel in these pooled sums, so even large but inconsistently directed typed-distance gaps collapse to a negligible global correction. This collapse is precisely why a pooled score cannot see sample-level structure, and motivates the per-sample construction of `CRoMa` (Section 3.2).

Both effects are visible on Camelyon. At the operating median k, every one of the 20 foundation models has fewer than half of all samples contributing to `RI`/`MaRI`, and more than half of them have fewer than one quarter (Table 3). Thus, the margin correction rides on a thin, model-specific slice of the data. Within that subset, the aggregate **SO**-versus-**OS** distance asymmetry is mild and inconsistently directed, leaving little room for `MaRI` to depart from `RI`. On Camelyon these two effects are confounded — thin support and a genuinely undirected asymmetry both predict a small Δ — so we defer the disambiguation to Section 3.4, where benchmarks with considerably larger support leave Δ just as small and pin the near-equivalence on the pooled design rather than on low support.

This failure mode is not incidental: it's a direct consequence of how the neighbourhood scale

is chosen. The operating k is derived from biological k-nearest-neighbour accuracy, computed over the full dataset. Same-biology, same-confounder (**SS**) neighbours often dominate the criterion used to select the scale. Yet these neighbours are discarded by `RI` and `MaRI`, which score only the **SO**/**OS** contrast. The selected neighbourhood can consequently be well matched to local biological retrieval while remaining too small to expose the typed evidence needed for robustness scoring. This mismatch is visible in the neighbour-rank structure. On Camelyon, even the first typed neighbour typically appears far beyond the evaluated neighbourhood: pooled across all 20 foundation models, the median rank at which at least one **SO** or **OS** neighbour appears is $\approx$ 149 among 20,400 candidates (Supplementary Table 12), whereas the evaluated neighbourhood contains only $k = 11$ neighbours. The full biology-versus-confounder *contrast* – both a nearest **SO** *and* a nearest **OS** – lies deeper still. For many anchors, the fixed neighbourhood therefore sits deep inside the **SS** shell, before even a single typed neighbour – let alone the contrast required by `RI` and `MaRI` – has appeared.

`MaRI`'s margin-awareness can refine the score where typed evidence is visible, but it cannot recover anchors excluded by construction. This exposes the central limitation of fixed-k robustness metrics: they condition evaluation on the appearance of typed neighbours within a pre-specified neighbourhood. As a result, pooled scores reflect a model-dependent subset of the data, complicating cross-model comparisons. A principled robustness metric requires a different construction: one that measures the signed distance margin between cross-confounder matches and same-confounder distractors directly, rather than conditioning on their appearance within a fixed neighbourhood.

3.2 `CRoMa` defines a typed margin beyond fixed neighbourhoods

The limitations of `RI` and `MaRI` arise from the same design choice: robustness is inferred from the composition of a fixed-k neighbourhood. `CRoMa` removes this bottleneck by recasting robustness as a direct distance comparison. For each sample, it asks whether cross-confounder biological matches (**SO**) lie closer than same-confounder biological distractors (**OS**). Rather than restricting the contrast to typed neighbours that happen to fall inside a fixed-k window, `CRoMa` searches outward to the m nearest neighbours of each type and compares their mean distances (Methods).

This construction yields a signed margin bounded to $(-1, 1)$, positive when biological matches across confounders are closer than confounder-matched distractors, and negative otherwise. `CRoMa` is defined for every sample whenever the required typed reference sets exist in the dataset — that is, when each biological class co-occurs with at least m confounders, and each confounder with at least m biological classes. Score definition is therefore governed by benchmark construction rather than by model-specific neighbourhood geometry. `CRoMa` consequently evaluates all encoders on the same eligible population, instead of letting each model determine which samples enter its pooled estimate.

All encoders strongly encode both biological class and acquisition centre. On Camelyon, biological k-NN accuracy is near-ceiling ($0.93 - 0.99$), while acquisition centre is almost perfectly decodable ($0.92 - 1$) across all 20 foundation models. What separates a robust representation from a shortcut-prone one is therefore not whether each signal is present, but which of the two dominates when they compete locally. We quantify that competition with the median per-sample `CRoMa` margin, which asks whether the typical sample lies closer to cross-confounder biological matches or to same-confounder biological distractors. Model-level medians range from -0.44 to 0.20 (Table 3) but do not form a smooth continuum. `CRoMa` separates models into three distinct regimes. At the top, a biology-dominant group includes `Virchow2` and `CONCH` (both 0.20), `GenBio-PathFM` and `CONCHv1.5` (both 0.19), followed closely by `H0-mini` (0.17) and `Virchow` (0.16). A middle band

lies close to zero, where biological and confounder structure are nearly balanced: `Midnight-12k` (0.11), `H-optimus-1` (0.08), `H-optimus-0` (0.05), `UNI2-h` and `MUSK` (both 0.04), `mSTAR` (0.02) and `Prov-GigaPath` (0.01). Finally, a confounder-dominant tail — 7 of the 20 encoders — drops sharply, from `UNI` (-0.03), `Hibou-B` (-0.09) and `GPFM` (-0.10) through `Phikon` and `Phikon-v2` (near -0.20) to `Prost40M` (-0.32) and `Hibou-L` (-0.44).

Where fixed-k metrics are defined, their rankings broadly agree with `CRoMa` on Camelyon ($\rho = 0.95$ versus `RI`; $\rho = 0.94$ versus `MaRI`). This agreement is reassuring: `CRoMa` does not overturn the fixed-k signal on anchors for which typed evidence is visible. Nevertheless, the metrics are not interchangeable. `H-optimus-1` and `Virchow` sit close under `MaRI` (0.677 versus 0.708), yet separate clearly under `CRoMa` (0.08 versus 0.16; Table 3). Conversely, `UNI2-h` and `Prov-GigaPath` have nearly identical `CRoMa` medians (0.04 versus 0.01), but differ substantially under `MaRI` (0.548 versus 0.369). These discrepancies reflect the different estimands: `RI` and `MaRI` summarize typed evidence only within a fixed neighbourhood and only on supported anchors, whereas `CRoMa` measures the signed typed margin over the full evaluation cohort.

Biological accuracy explains the spread of `CRoMa` values only weakly ($\rho = 0.56$ against biological k-NN accuracy): `CONCH`, `CONCHv1.5`, `Hibou-B`, and `Hibou-L` all recover biology near-identically (k-NN accuracy ≈ 0.97), yet span almost the entire `CRoMa` range. What explains it instead is the confounder accuracy. Across the 20 pathology encoders, the balanced accuracy with which a k-NN probe recovers the acquisition centre rank-predicts the `CRoMa` median almost perfectly ($\rho = -0.93$): the more decodable the centre, the more confounder-dominant the margin. The same probe predicts `RI` and `MaRI` at least as well ($\rho = -0.95$ and -0.94, respectively), so as a device for *ranking* encoders no pooled score here recovers much that a centre probe does not. What a scalar probe cannot express, however, is the shape of the margin distribution below its median: where the lower tail decouples from the centre it also decouples from the probe — on TCGA-4x4, the benchmark on which the probe is *furthest* from ceiling ($[0.49, 0.79]$, against a chance level of 0.25), it rank-predicts $\texttt{LTM}_{10}$ at only $\rho = -0.21$, far below its near-perfect hold on the pooled scores. That gap is where the per-sample construction earns its keep, and we turn to it next (Section 3.3).

3.3 Per-sample margins expose hidden fragility

A pooled median margin asks whether the typical sample is biology-dominant, but it does not reveal how *frequently*, nor how *strongly*, samples fall on the confounder-dominant side of the distribution. This distinction is important for clinical deployment, where risk may be concentrated in a vulnerable subset even when the centre of the distribution appears robust. Because `CRoMa` is defined at the sample level, it yields a cohort-wide distribution rather than a single aggregate score. This distribution constitutes the model's robustness fingerprint on a given cohort. We characterise its vulnerable tail using two complementary quantities: *failure prevalence*, the confounder-dominant fraction $F(0)$, and *failure severity*, the mean of the worst decile, $\texttt{LTM}_{10}$ (Methods). Together, these distinguish typical model robustness from the size and severity of its vulnerable tail. Both are compact summaries chosen to facilitate comparison across models: the full distribution remains the primary readout.

On Camelyon, no representation is uniformly robust. Every one of the 20 pathology encoders retain a confounder-dominated lower tail: all have $\texttt{LTM}_{10} < 0$ (Table 3). The prevalence of confounder-dominated samples, however, varies widely: from 12.9% for `Virchow2`, the leading encoder by median margin, to 99.3% for `Hibou-L`. Median performance, failure prevalence and tail severity are

Model	bio bacc	conf bacc	RI	MaRI	Δ	CRoMa	$F(0)$	LTM_{10}	support
Virchow2	**0.988**	0.958	0.806	0.823	+0.017	**0.20**	0.129	-0.11	31.4%
CONCH	0.971	0.956	0.662	0.626	−0.037	0.20	0.225	-0.20	35.9%
GenBio-PathFM	0.983	0.928	**0.842**	**0.850**	+0.008	0.19	**0.092**	**-0.07**	38.3%
CONCHv1.5	0.971	0.915	0.774	0.763	−0.012	0.19	0.174	-0.14	**46.3%**
H0-mini	0.969	0.927	0.741	0.718	−0.023	0.17	0.180	-0.16	38.7%
Virchow	0.980	0.946	0.751	0.708	−0.043	0.16	0.221	-0.18	26.4%
Midnight-12k	0.976	0.984	0.478	0.408	−0.070	0.11	0.354	-0.35	19.8%
H-optimus-1	0.986	0.978	0.664	0.677	+0.013	0.08	0.219	-0.14	17.2%
H-optimus-0	0.982	0.966	0.659	0.652	−0.007	0.05	0.315	-0.15	23.7%
UNI2-h	0.986	0.987	0.515	0.548	+0.033	0.04	0.370	-0.21	13.0%
MUSK	0.958	0.983	0.366	0.297	−0.069	0.04	0.403	-0.22	28.0%
mSTAR	0.979	0.984	0.460	0.434	−0.026	0.02	0.418	-0.18	18.0%
Prov-GigaPath	0.979	0.991	0.375	0.369	−0.007	0.01	0.470	-0.19	14.4%
UNI	0.982	0.999	0.108	0.092	−0.015	-0.03	0.651	-0.22	9.6%
Hibou-B	0.973	0.999	0.057	0.041	−0.017	-0.09	0.737	-0.36	13.4%
GPFM	0.955	0.999	0.034	0.017	−0.017	-0.10	0.753	-0.36	20.9%
Phikon	0.955	1.000	0.009	0.004	−0.005	-0.20	0.905	-0.48	16.6%
Phikon-v2	0.954	1.000	0.019	0.008	−0.011	-0.21	0.932	-0.50	16.9%
Prost40M	0.926	1.000	0.015	0.002	−0.012	-0.32	0.922	-0.64	27.2%
Hibou-L	0.971	1.000	0.013	0.001	−0.011	-0.44	0.993	-0.66	12.1%
DINOv2-B	0.919	0.912	0.561	0.507	−0.053	0.05	0.345	-0.18	68.0%

Table 3: **Representation robustness on Camelyon.** Pooled results for 20 tile-level pathology foundation models, ordered by median CRoMa (m=5), with the natural-image control DINOv2-B shown separately. All models are evaluated at the shared operating point k=11, the dataset median of the per-model biological $k^\star$. Columns: biological and confounder k-NN balanced accuracy (bio bacc and conf bacc; confounder: medical centre); pooled RI and MaRI; Δ=MaRI−RI; median CRoMa; $F(0)$, the fraction with CRoMa < 0; LTM_{10}, the mean of the lowest decile; and support, the fraction of samples effectively contributing to RI/MaRI. Bold denotes the best value in each score column (conf bacc and Δ are diagnostics).

not interchangeable. Although CONCH matches Virchow2 at the median (0.20), it fails more often ($F(0) = 22.5\%$ versus 12.9%) and more severely ($\text{LTM}_{10} = -0.20$ versus -0.11). Conversely, CONCH fails less often than H-optimus-0 (22.5% versus 31.5%), but more severely (-0.20 versus -0.15). No single pooled score can therefore capture all three properties. For model selection, median margin and tail severity define distinct axes of robustness, making encoder selection a Pareto problem (Figure 3). Plotting encoders in this plane exposes the trade-off between typical and worst-tail robustness, while failure prevalence provides a complementary measure of how broadly vulnerability is distributed across the cohort. An encoder is dominated if another performs at least as well on both axes and strictly better on one. Those that survive form the Pareto frontier. On Camelyon, this frontier contains only two of the 20 pathology encoders: Virchow2 and GenBio-PathFM. Choosing a representation is therefore a trade-off between typical robustness and tail severity rather than the maximisation of a single score, and only a per-sample construction makes both axes measurable at once.

These summaries remain scalar projections, and the full margin distribution retains structure that none captures. For example, Midnight-12k and H-optimus-1 have markedly different distributional shapes (Figure 4). The margins of H-optimus-1 are tightly concentrated around the median, whereas those of Midnight-12k span nearly the full range. The per-sample distribution also enables finer-grained analysis: distinct modes may identify subpopulations governed by different

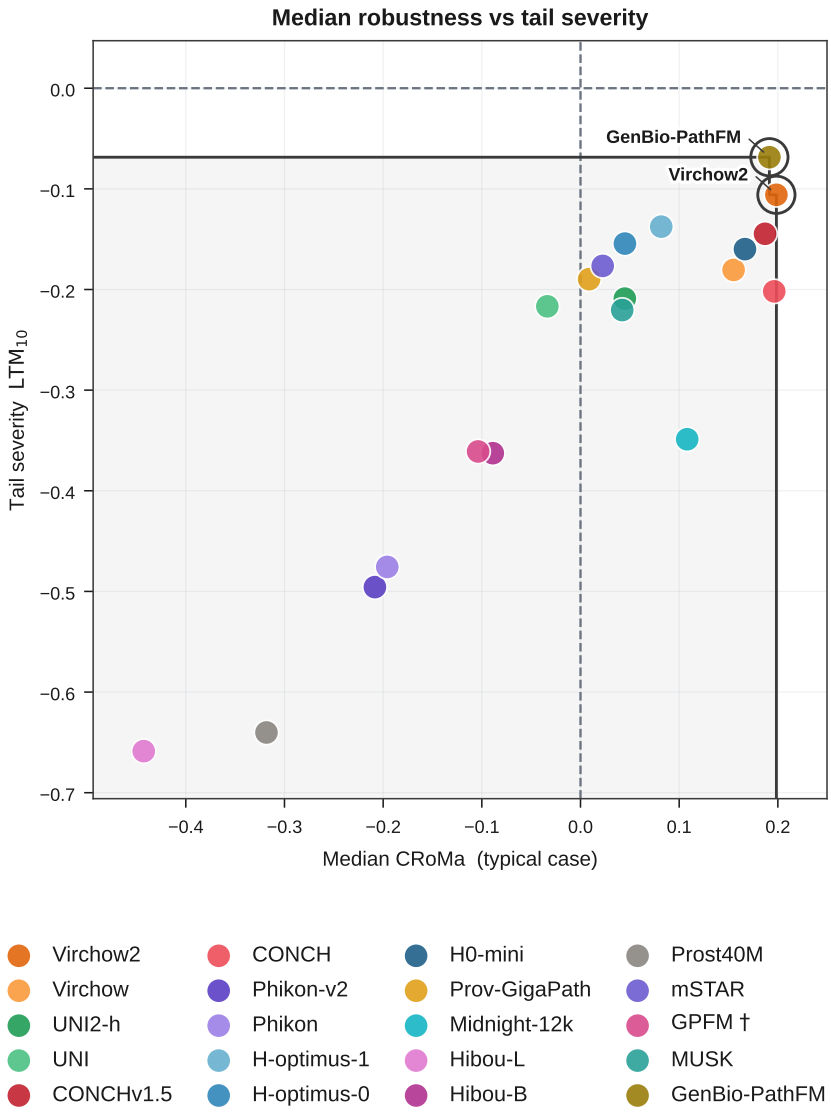


Figure 3: **Median and lower-tail `CRoMa` on Camelyon.** Median `CRoMa` is plotted against worst-decile mean $\texttt{LTM}_{10}$ for the 20 pathology encoders evaluated in this study. Higher values are preferable on both axes. Ringed points form the upper-right Pareto frontier, while shaded points are dominated on both axes. † marks the Camelyon-exposed encoder (`GPFM`).

margin regimes, and linkage to the source samples permits stratification by acquisition centre or biological class to localise fragility.

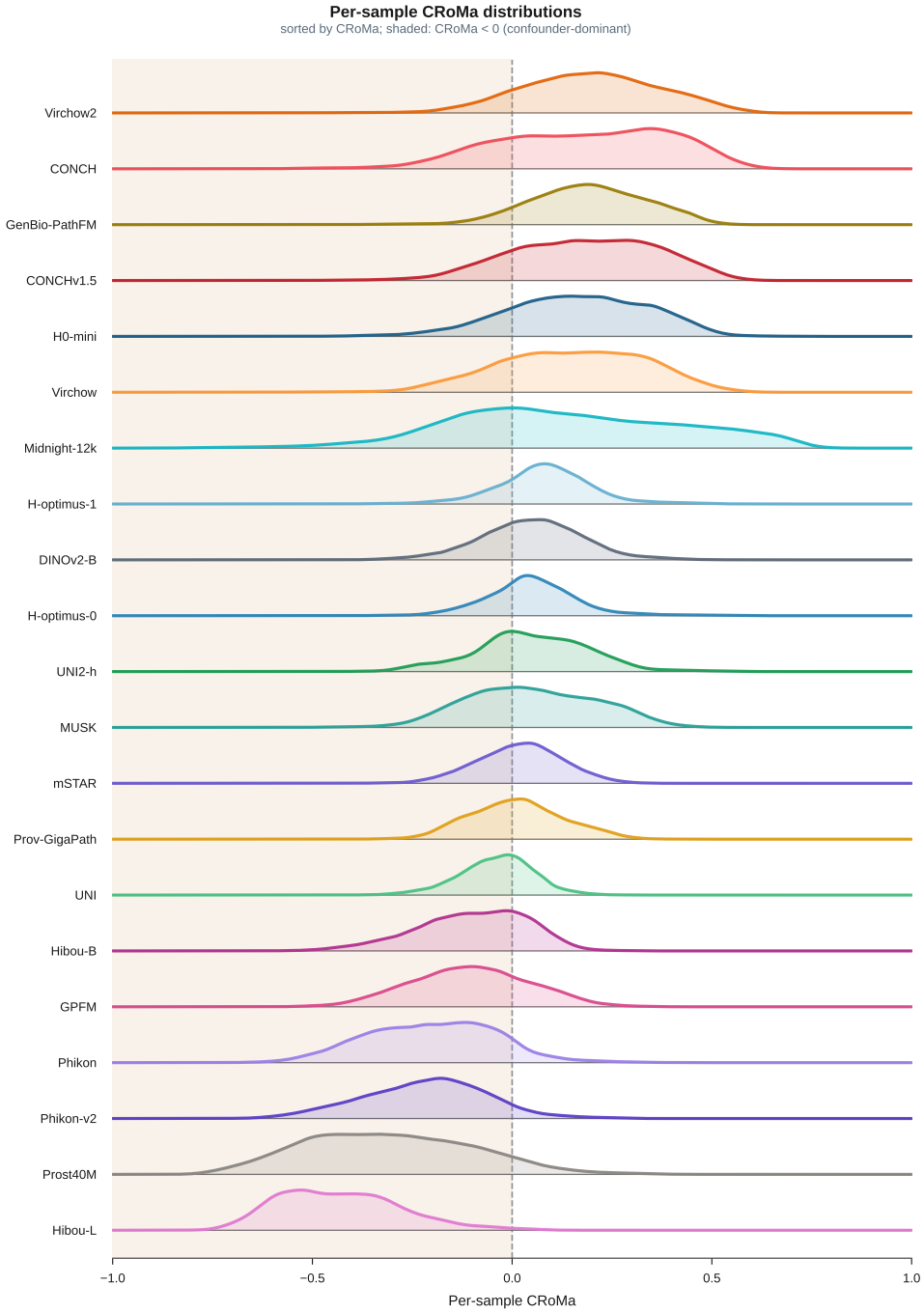


Figure 4: **Per-sample `CRoMa` distributions on Camelyon.** Ridgeline distributions for the 20 pathology encoders and the natural-image control (`DINOv2-B`), ordered by median `CRoMa`. The dashed line denotes `CRoMa` $= 0$, while shading denotes the confounder-dominant region (`CRoMa` < 0).

3.4 Robustness insights are consistent across benchmarks

Having used Camelyon as a stringent test of fixed-neighbourhood robustness, we next asked whether `CRoMa` captures a reproducible representation-level signal across benchmarks. We evaluated the same encoders on TCGA-4x4 (Table 4) and Tolkach-ESCA (Table 5). For completeness, we also report PathoROB's TCGA-2x2 configuration in Supplementary Section A.11.

Both datasets provide substantially higher `RI`/`MaRI` support: $99\% - 100\%$ on TCGA-4x4 and $67\% - 100\%$ on Tolkach-ESCA (against $10\% - 46\%$ on Camelyon). This is a direct consequence of the larger operating point these cohorts select (k=71 and 61 respectively, against k=11 on Camelyon): a wider fixed neighbourhood is more likely to contain a typed **SO**/**OS** contrast. This higher support lets us revisit the count-versus-margin comparison of Section 3.1 on a far thicker slice of each cohort and settle the confound left open there. On Camelyon, `MaRI` barely departed from `RI` (Δ = `MaRI` − `RI` within ± 0.07), but low support left the cause ambiguous: the near-equivalence could reflect a genuinely undirected typed-distance asymmetry, which the pooled index collapses, or merely the thin slice on which the pooled margin was evaluated. Were thin support the explanation, the correction should grow once `RI` and `MaRI` become defined on most of the cohort. It does not: Δ stays within ± 0.04 on TCGA-4x4 and ± 0.04 on Tolkach-ESCA across all 20 pathology encoders. The near-equivalence of pooled counts and pooled margins (Spearman ρ =0.99 and 0.98, respectively) is therefore intrinsic to the fixed-k, pooled design, which is precisely what motivates the per-sample construction of `CRoMa`. Where fixed-k scores are defined, `CRoMa` again broadly tracks their rankings (ρ =0.95 and 0.93 versus `RI`, 0.95 and 0.92 versus `MaRI` on TCGA-4x4 and Tolkach-ESCA), matching the agreement seen on Camelyon (ρ =0.95 and 0.94; all nine within-benchmark rank correlations are collected in Table 9).

Across tile-level benchmarks, `CRoMa` identifies consistent robustness patterns while also reflecting differences in benchmark difficulty. In contrast with Camelyon, where several encoders showed negative or near-zero median margins, TCGA-4x4 and Tolkach-ESCA shifted the margin distributions upward, with `CRoMa` spanning $[-0.01, 0.40]$ and $[0.11, 0.58]$ respectively. At the median, only one of the 20 encoders remains confounder-dominant on TCGA-4x4, and none on Tolkach-ESCA. Howver, this upward shift did not eliminate the failure modes observed on Camelyon: even when the typical sample was robust, the worst decile remained confounder-dominant for every encoder ($\texttt{LTM}_{10} < 0$; Tables 4 and 5). The per-sample distributions make this visible: on both benchmarks every encoder keeps mass on the confounder-dominant side of the boundary (Supplementary Figures 10 and 11). Model rankings were largely stable across datasets (pairwise Spearman $\rho \in [0.88, 0.92]$; Table 9). Encoders ranked among the most robust on Camelyon generally remained competitive on TCGA-4x4 and Tolkach-ESCA, whereas low-margin models tended to remain lower ranked. The `CONCH` family and `GenBio-PathFM` were consistently among the top five, with `H0-mini` close behind on all three and `Virchow2` leading Camelyon but slipping to fifth and seventh elsewhere. `Prost40M`, `Hibou-B`, and the `Phikon` models remained near the bottom.

The main exception was `Midnight-12k`, which ranked seventh on Camelyon (`CRoMa`= 0.11) but first on both TCGA-4x4 (0.40) and Tolkach-ESCA (0.58). Because `Midnight-12k` was pretrained exclusively on TCGA, its performance on TCGA-4x4 may partly reflect pretraining–benchmark overlap. A within-cohort provenance analysis supports this interpretation but also shows that such overlap is not sufficient to explain robustness (Supplementary Section A.7). Other TCGA-pretrained models did not show the same advantage on TCGA-4x4: `Phikon` remained weak overall, whereas `Phikon-v2` exhibited no comparable rank shift, indicating that provenance overlap alone

is not sufficient to guarantee higher CRoMa. Instead, robustness appears to reflect an interaction between pretraining provenance, data scale, training recipe, and the intrinsic complexity of the evaluated samples.

Because median performance and tail severity capture distinct robustness properties, cross-benchmark consistency is best assessed in the median–tail plane rather than through a single ranking. Supplementary Figures 14 and 15 show that the pattern observed on Camelyon extends to TCGA-4x4 and Tolkach-ESCA: the encoder with the highest median margin does not have the least severe worst decile, no encoder dominates the plane, and only a small subset lies on the Pareto frontier. Combining the three tile-level benchmarks requires a scale-free summary because the dispersion of raw CRoMa margins differs several-fold across datasets. Averaging model medians would therefore overweight benchmarks with broader score distributions and amplify the in-distribution advantage described above. We avoid both distortions by ranking encoders within each benchmark and averaging their ranks. Figure 5 plots mean median-CRoMa rank against mean LTM_{10} rank. Across benchmarks, only four of the 20 encoders remain undominated – CONCH, GenBio-PathFM, CONCHv1.5 and H-optimus-1 – and none leads on both axes.

These conclusions extend to whole-slide representations. On the PCaBio slide-level benchmark, CRoMa again separates biological discriminability from confounder invariance (Supplementary Section A.12). The confounder is near-perfectly decodable for all four whole-slide encoders, yet median margins span $[-0.41, 0.26]$, and only PRISM is biology-dominant (CRoMa=0.26). Per-sample summaries remain non-redundant: PRISM and MOOZY differ sharply in median margin and failure prevalence but have near-identical worst-decile severity ($\text{LTM}_{10} = -0.39$ versus -0.41).

Together, these results show CRoMa captures a reproducible representation-level robustness signal across benchmarks and scales while retaining sensitivity to vulnerable subpopulations. We next asked whether it can predict the downstream failure mode it is designed to expose: do lower margins identify models more susceptible to shortcut learning?

3.5 CRoMa predicts downstream shortcut susceptibility

The value of a representation-level robustness metric ultimately lies in its ability to anticipate downstream failures under supervised adaptation. We therefore replicate the shortcut-learning protocol introduced in [19]: linear probes are trained to predict the biological label from frozen embeddings while the training set is progressively biased by increasingly introducing spurious confounder–biology correlation, quantified by Cramér's V. A model that entangles confounders with biology in representational space hands the probe an exploitable shortcut and should therefore suffer a larger drop precisely when that shortcut is removed. Performance is thus evaluated on balanced test sets in which this shortcut no longer supports the prediction: an in-domain setting using confounder labels seen during training, and an out-of-domain setting using held-out confounder labels. APD averages the relative performance drop across the biased training splits with respect to the balanced baseline ($V = 0$): values near zero indicate stable generalization, whereas increasingly negative values indicate stronger shortcut susceptibility. Because its out-of-domain arm requires held-out confounders, APD is measured on a broader row set than the geometry metrics: it reads each model's frozen embeddings over the full in-domain *and* out-of-domain data pool of every benchmark (Table 6), whereas RI, MaRI and CRoMa use PathoROB's robustness-evaluation view (Table 1).

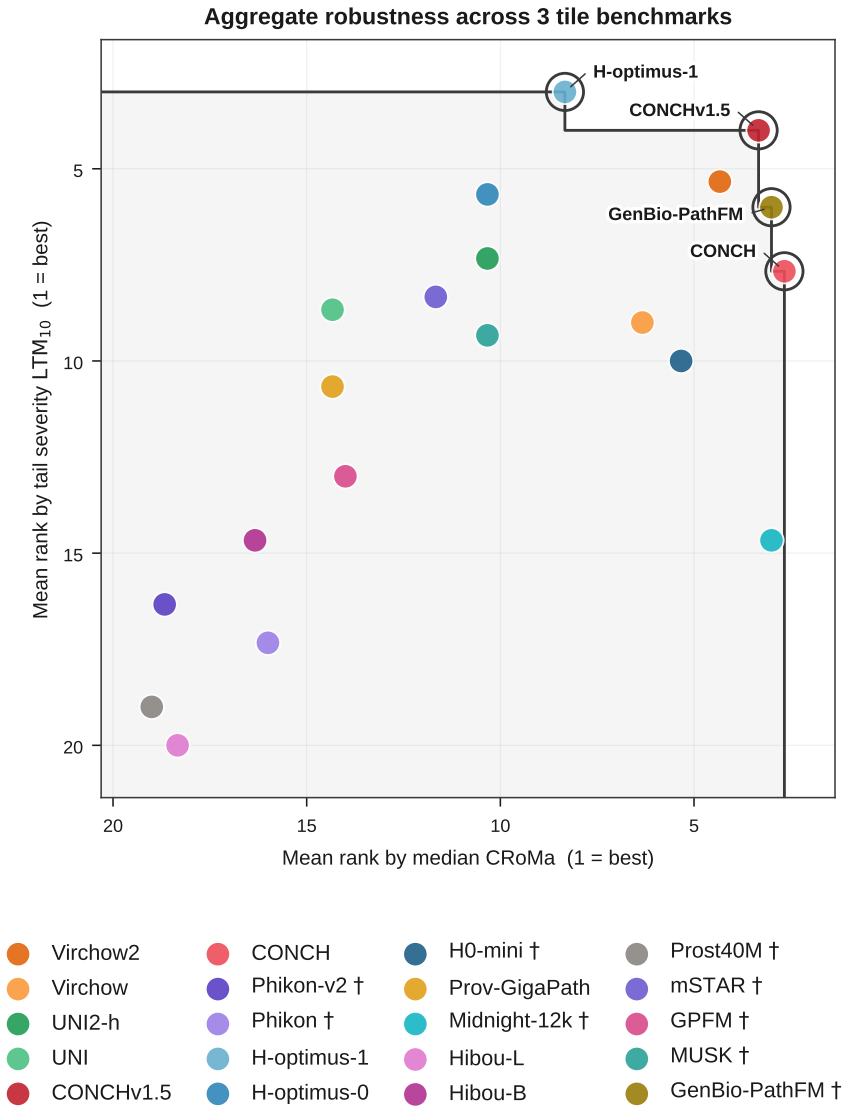


Figure 5: **Median and lower-tail `CRoMa` ranks across tile benchmarks.** Each of the 3 tile benchmarks ranks the 20 pathology encoders by median `CRoMa` (rank 1 = highest median) and by worst-decile mean $\texttt{LTM}_{10}$ (rank 1 = mildest tail). The axes show the mean rank across benchmarks, with lower (=better) ranks plotted toward the upper-right corner. Ringed points form the Pareto frontier, while shaded points are dominated on both axes. † marks the 9 TCGA-exposed encoders.

Model	bio bacc	conf bacc	RI	MaRI	Δ	CRoMa	$F(0)$	LTM_{10}	support
Midnight-12k†	**0.882**	0.559	**0.898**	**0.936**	+0.037	**0.40**	**0.140**	-0.21	99.4%
GenBio-PathFM†	0.851	0.695	0.757	0.780	+0.023	0.16	0.242	-0.19	99.9%
CONCHv1.5	0.811	0.492	0.828	0.851	+0.024	0.15	0.193	-0.13	100.0%
CONCH	0.790	0.487	0.801	0.825	+0.024	0.15	0.216	-0.15	**100.0%**
Virchow2	0.825	0.594	0.771	0.795	+0.025	0.13	0.239	-0.17	**100.0%**
H0-mini†	0.820	0.617	0.737	0.754	+0.017	0.12	0.257	-0.19	99.9%
Virchow	0.789	0.656	0.697	0.706	+0.009	0.09	0.301	-0.18	**100.0%**
H-optimus-1	0.878	0.676	0.775	0.786	+0.012	0.09	0.220	**-0.10**	99.8%
UNI2-h	0.847	0.737	0.711	0.725	+0.013	0.07	0.265	-0.12	99.3%
mSTAR†	0.817	0.690	0.696	0.703	+0.007	0.06	0.298	-0.12	99.9%
H-optimus-0	0.846	0.692	0.707	0.711	+0.004	0.05	0.287	-0.11	99.9%
MUSK†	0.724	0.589	0.674	0.680	+0.006	0.05	0.332	-0.14	**100.0%**
Prov-GigaPath	0.831	0.672	0.683	0.679	−0.005	0.05	0.308	-0.15	99.9%
UNI	0.807	0.712	0.673	0.678	+0.005	0.05	0.320	-0.12	100.0%
Hibou-L	0.759	0.737	0.577	0.541	−0.037	0.04	0.405	-0.25	99.7%
GPFM†	0.759	0.725	0.612	0.604	−0.008	0.04	0.387	-0.15	100.0%
Phikon†	0.792	0.788	0.577	0.577	−0.000	0.04	0.396	-0.20	99.2%
Hibou-B	0.768	0.722	0.600	0.590	−0.010	0.04	0.387	-0.18	99.9%
Phikon-v2†	0.790	0.782	0.550	0.544	−0.006	0.03	0.409	-0.19	99.4%
Prost40M†	0.635	0.718	0.464	0.459	−0.005	-0.01	0.528	-0.25	**100.0%**
DINOv2-B	0.607	0.580	0.540	0.513	−0.028	0.01	0.468	-0.12	100.0%

Table 4: **Representation robustness on TCGA-4x4.** Pooled results for the 20 tile-level pathology foundation models, ordered by median CRoMa (m=5), with the natural-image control DINOv2-B shown separately. All models are evaluated at the shared operating point k=71, the dataset median of the per-model biological $k^\star$. Columns: biological and confounder k-NN balanced accuracy (bio bacc and conf bacc; confounder: medical centre); pooled RI and MaRI; Δ=MaRI − RI; median CRoMa; $F(0)$, the fraction with CRoMa < 0; LTM_{10}, the mean of the lowest decile; and support, the fraction of samples effectively contributing to RI/MaRI. Bold denotes the best value in each score column (conf bacc and Δ are diagnostics). † marks the 9 TCGA-exposed encoders (Table 2).

Across the 20 tile-level pathology encoders, CRoMa consistently tracks this downstream vulnerability. Models with biology-dominant margins incurred smaller in-domain drops, whereas models with confounder-dominant margins showed the largest shortcut-induced degradation (APD_{ID}: Spearman $\rho = 0.84$ pooled over 60 model–benchmark pairs). The association was weaker out of domain, but remained directionally consistent (APD_{OOD}: pooled Spearman $\rho = 0.78$), as expected for an evaluation that combines shortcut reliance with transfer to unseen confounder labels. We show the full relationship for Camelyon (Fig. 6). The same pattern holds on TCGA-4x4 and Tolkach-ESCA (Supplementary Figures 12 and 13), confirming CRoMa captures more than a geometric property of frozen embeddings: it identifies representations that are more likely to support confounder-based shortcuts once a downstream predictor is trained. Unlike APD, which requires repeated supervised training under engineered confounder–label correlations, CRoMa provides an upstream readout of shortcut susceptibility directly from the representation space.

Model	bio bacc	conf bacc	RI	MaRI	Δ	CRoMa	$F(0)$	LTM_{10}	support
Midnight-12k	0.976	0.728	0.943	0.941	−0.002	**0.58**	0.051	-0.08	99.0%
CONCH	0.973	0.654	0.951	0.957	+0.006	0.44	0.045	-0.04	99.8%
CONCHv1.5	0.973	0.633	0.952	0.964	+0.012	0.39	0.043	-0.03	**99.9%**
GenBio-PathFM	**0.981**	0.598	**0.960**	**0.964**	+0.004	0.39	**0.038**	**-0.02**	**99.9%**
H0-mini	0.967	0.642	0.935	0.946	+0.011	0.38	0.058	-0.07	99.7%
Virchow	0.970	0.703	0.935	0.943	+0.008	0.37	0.053	-0.05	99.6%
Virchow2	0.978	0.613	0.954	0.957	+0.002	0.35	0.040	-0.04	99.6%
MUSK	0.969	0.739	0.924	0.931	+0.008	0.29	0.063	-0.07	99.2%
H-optimus-1	0.977	0.680	0.940	0.949	+0.008	0.26	0.049	-0.04	99.1%
GPFM	0.969	0.841	0.883	0.902	+0.018	0.24	0.095	-0.10	97.6%
H-optimus-0	0.971	0.731	0.911	0.919	+0.007	0.23	0.076	-0.08	98.5%
UNI2-h	0.976	0.767	0.916	0.927	+0.012	0.22	0.064	-0.06	97.7%
mSTAR	0.970	0.818	0.881	0.898	+0.017	0.19	0.087	-0.09	97.7%
Phikon	0.962	0.898	0.772	0.782	+0.010	0.17	0.179	-0.19	81.5%
UNI	0.973	0.835	0.880	0.891	+0.011	0.17	0.086	-0.08	95.3%
Hibou-B	0.967	0.916	0.774	0.775	+0.000	0.13	0.188	-0.17	85.7%
Prov-GigaPath	0.962	0.929	0.707	0.746	+0.039	0.13	0.236	-0.16	79.1%
Prost40M	0.911	0.838	0.701	0.721	+0.019	0.13	0.277	-0.24	96.2%
Phikon-v2	0.956	0.896	0.741	0.746	+0.005	0.12	0.225	-0.17	83.3%
Hibou-L	0.955	0.960	0.624	0.586	−0.038	0.11	0.315	-0.29	67.0%
DINOv2-B	0.905	0.535	0.876	0.867	−0.008	0.18	0.099	-0.07	100.0%

Table 5: **Representation robustness on Tolkach-ESCA.** Pooled results for the 20 tile-level pathology foundation models, ordered by median CRoMa (m=5), with the natural-image control DINOv2-B shown separately. All models are evaluated at the shared operating point k=61, the dataset median of the per-model biological $k^\star$. Columns: biological and confounder k-NN balanced accuracy (bio bacc and conf bacc; confounder: medical centre); pooled RI and MaRI; Δ=MaRI − RI; median CRoMa; $F(0)$, the fraction with CRoMa < 0; LTM_{10}, the mean of the lowest decile; and support, the fraction of samples effectively contributing to RI/MaRI. Bold denotes the best value in each score column (conf bacc and Δ are diagnostics).

	Robustness metrics	APD (downstream probe)		
Benchmark	(RI/MaRI/CRoMa)	In-domain	Out-of-domain	Full pool
Camelyon	20,400 (2)	20,400 (2)	2,002 (3)	22,402 (5)
TCGA-4x4	5,760 (4)	5,760 (4)	2,400 (4)	8,160 (8)
Tolkach-ESCA	9,000 (3)	10,800 (2)	5,500 (2)	16,300 (4)

Table 6: **Robustness metrics and APD use different evaluation rows.** Entries give tile counts, with the number of contributing centres or cohorts in parentheses. RI, MaRI and CRoMa use the PathoROB robustness view. APD uses in-domain and held-out out-of-domain pools from the same frozen encoders. For Camelyon and TCGA-4 × 4, the robustness rows equal the in-domain rows and APD adds held-out centres. For Tolkach-ESCA, the robustness view (UKK, WNS and CHA) and in-domain APD view (WNS and CHA) are distinc. UKK is reused out of domain.

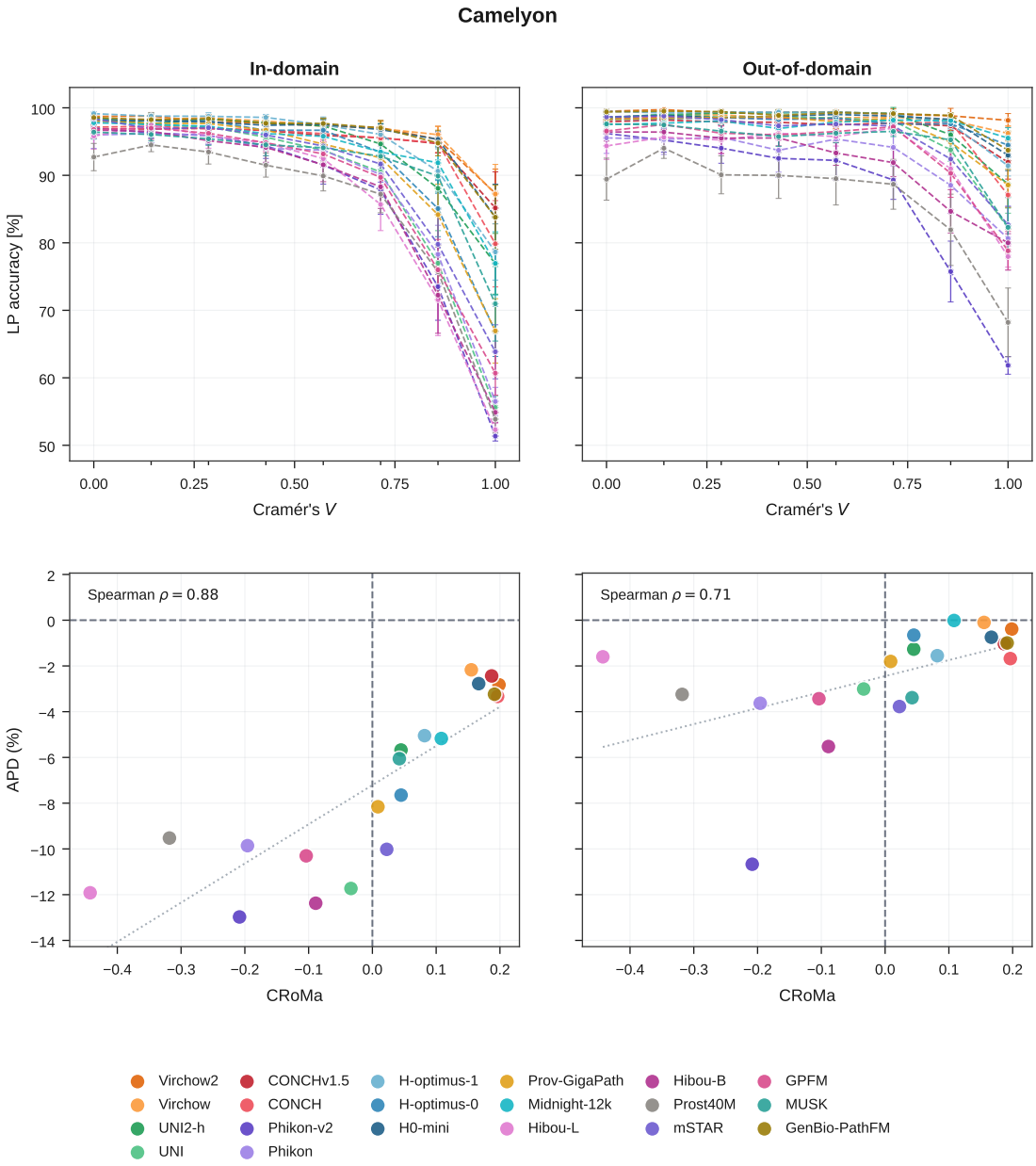


Figure 6: **`CRoMa` against `APD` on Camelyon.** Columns show in-domain evaluation (left, test centres held fixed) and out-of-domain evaluation (right, unseen centres). **Top row:** linear-probe accuracyon the balanced test set as the centre–biology correlation in the training set increases from V=0 to V=1. Dashed curves denote the 20 tile-level encoders and error bars are 95% confidence intervals over repeated seeding. **Bottom row:** pooled `CRoMa`(m=5) versus `APD`, one point per encoder. Dashed lines mark `CRoMa` = 0 and `APD` = 0.

4 Discussion

The value of a pathology foundation model lies in its ability to capture biologically meaningful signals while remaining robust to irrelevant non-biological variation. Biological discriminability and robustness are complementary dimensions of representation quality: a useful representation should support accurate downstream prediction without allowing batch effects from tissue preparation, staining or scanning to dominate its geometry. Strong biological encoding is therefore insufficient when non-biological cues provide an equally accessible shortcut for prediction. In this study, we show that existing fixed-neighbourhood scores capture a meaningful robustness signal but incompletely characterise its geometry and distribution across samples. By recasting robustness as a signed distance margin over the full evaluation cohort, `CRoMa` exposes confounder-dominated pockets hidden by pooled averages and anticipates downstream shortcut susceptibility.

The typed contrast introduced by `RI` is conceptually well founded: cross-confounder biological matches signal strong biological structure, whereas same-confounder biological distractors indicate that non-biological cues are shaping local geometry. Yet `RI` reduces this contrast to counts and is therefore blind to margins. `MaRI` restores geometric information through distance weighting. Its close agreement with `RI`, however, exposes a deeper limitation of the fixed-k, pooled design. A sample contributes only if its pre-specified neighbourhood contains at least one of the neighbour types entering the score. Otherwise, it is omitted in a model-dependent manner, undermining cross-model comparability. Pooling over the remaining samples further allows opposing sample-level asymmetries to cancel. A principled robustness measure therefore requires a different construction: one that defines the contrast for every sample, preserves its direction and magnitude, and treats aggregate scores as views of a distribution rather than as the estimand itself.

Treating robustness as a distribution reveals what no pooled aggregate can: no foundation model is uniformly robust. Even models with biology-dominant median margins retain confounder-dominated samples, and every evaluated tile encoder exhibits a negative lower tail. The resulting picture is not a division between robust and non-robust encoders, but a spectrum of robustness profiles in which central tendency, failure prevalence and failure severity vary partly independently. Models with similar median margins can therefore differ markedly in how often, and how strongly, non-biological variation outweighs biological signal: failures may be rare but pronounced, frequent but modest, or both frequent and severe. Model selection is therefore better framed as a Pareto problem balancing central tendency against tail severity, rather than driven by a single score. After aggregating encoder ranks across the three tile benchmarks, only a small subset remains undominated in the median–tail plane. None, however, leads on both axes.

One encoder nevertheless stood out for its consistency. `GenBio-PathFM` was the only model on the median–tail Pareto frontier in all three tile benchmarks, while remaining among the top four by median `CRoMa` in each. Its JEDI recipe adds JEPA-based refinement to a DINO-family pretraining stage [16], departing from the predominantly DINO/DINOv2-only recipes used by contemporary pathology encoders. Differences in model scale, data curation, corpus composition and pretraining–evaluation overlap prevent attributing this consistency specifically to JEPA. The result nonetheless motivates systematic investigation of hybrid and alternative pretraining objectives for learning more robust pathology foundation models.

A model that appears robust on average may still be unsuitable if its failures are concentrated in clinically important subsets. By preserving sample-level heterogeneity, `CRoMa` reframes robust-

ness from a pooled ranking statistic into a sample-resolved map of representational vulnerability. This finer resolution broadens the role of robustness assessment beyond model ranking: low-margin samples can be interrogated for enrichment of particular biological subclasses, institutions, or patient characteristics. Such enrichment would reveal structured failure modes concealed by an otherwise acceptable aggregate score. Identifying these pockets could guide targeted data collection, data augmentation and model pretraining, while motivating evaluation protocols that deliberately stress the conditions under which robustness is weakest. `CRoMa` therefore serves not only as a benchmark statistic, but also as a diagnostic tool for identifying where transferability is most at risk.

The broad agreement of `CRoMa` rankings across benchmarks suggests that it captures a reproducible property of learned representations, even as the distributions themselves shift with the biological task, confounder and cohort. Robustness is therefore neither purely model-intrinsic nor wholly dataset-specific: some encoders consistently privilege biological over technical structure, but the extent and location of their failures remain context-dependent. More importantly, `CRoMa` anticipated downstream shortcut susceptibility: models with more biology-dominant margins sustained smaller shortcut-induced performance losses. Computed directly from frozen embeddings, it provides a tractable proxy for transferability before a task-specific predictor is fitted. Comprehensive evaluation across institutions and population strata requires large, labelled and harmonised cohorts, together with repeated downstream training — resources that are often unavailable or prohibitively costly. `CRoMa` does not replace external validation, but enables a compact, deliberately structured multi-institutional cohort to identify representations most likely to fail under distribution shift and and prioritise the models, populations and conditions requiring deeper evaluation.

Taken together, these findings support a more stringent standard for robustness evaluation in pathology foundation models. By recasting robustness as a distributional property of representation space rather than a single pooled score, `CRoMa` measures not only whether a model is robust on average, but where, and by how much, confounders override biological similarity. Making this boundary explicit links representation geometry to shortcut susceptibility and directs validation towards the samples and settings in which transfer is most likely to fail. This sample-resolved view provides a more rigorous basis for assessing whether pathology foundation models retain biologically meaningful structure across institutions and populations. Future work should investigate pretraining strategies that prevent non-biological variation from organising representations without suppressing clinically relevant signal. Distributional robustness analyses can support this effort by identifying the biological subclasses, institutions and population subgroups concentrated in low-margin regions.

References

[1] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Ro-

hith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2022. arXiv:2108.07258.

[2] Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, et al. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nature Medicine*, 28(1):154–163, 2022.

[3] Gabriele Campanella, Shengjia Chen, Manbir Singh, Ruchika Verma, Silke Muehlstedt, Jennifer Zeng, Aryeh Stock, Matt Croken, Brandon Veremis, Abdulkadir Elmas, et al. A clinical benchmark of public self-supervised pathology foundation models. *Nature Communications*, 16(1):3640, 2025.

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021. arXiv:2104.14294.

[5] Richard J Chen, Tong Ding, Ming Y Lu, Drew F K Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L Weishaupt, Judy J Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nat. Med.*, 30(3):850–862, March 2024.

[6] Edwin D. de Jong, Eric Marcus, and Jonas Teuwen. Current pathology foundation models are unrobust to medical center differences, 2025. arXiv:2501.18055.

[7] Taher Dehkharghanian, Azam Asilian Bidgoli, Abtin Riasatian, Pooria Mazaheri, Clinton J V Campbell, Liron Pantanowitz, H R Tizhoosh, and Shahryar Rahnamayan. Biased data, biased AI: deep networks predict the acquisition site of TCGA images. *Diagn. Pathol.*, 18(1):67, May 2023.

[8] Tong Ding, Sophia J. Wagner, Andrew H. Song, Richard J. Chen, Ming Y. Lu, Andrew Zhang, Anurag J. Vaidya, Guillaume Jaume, Muhammad Shaban, Ahrong Kim, Drew F. K. Williamson, Harry Robertson, Bowen Chen, Cristina Almagro-Pérez, Paul Doucet, Sharifa Sahai, Chengkuan Chen, Christina S. Chen, Daisuke Komura, Akihiro Kawabe, Mieko Ochi, Shinya Sato, Tomoyuki Yokose, Yohei Miyagi, Shumpei Ishikawa, Georg Gerber, Tingying Peng, Long Phi Le, and Faisal Mahmood. A multimodal whole-slide foundation model for pathology. *Nature Medicine*, 31(11):3749–3761, 2025.

[9] Alexandre Filiot, Nicolas Dop, Oussama Tchita, Auriane Riou, Rémy Dubois, Thomas Peeters, Daria Valter, Marin Scalbert, Charlie Saillard, Geneviève Robin, and Antoine Olivier. Distilling foundation models for robust and efficient models in digital pathology. In *Medical Image*

Computing and Computer Assisted Intervention – MICCAI 2025, volume 15966 of *Lecture Notes in Computer Science*, pages 162–172. Springer Nature Switzerland, 2025.

[10] Alexandre Filiot, Ridouane Ghermi, Antoine Olivier, Paul Jacob, Lucas Fidon, Alice Mac Kain, Charlie Saillard, and Jean-Baptiste Schiratti. Scaling self-supervised learning for histopathology with masked image modeling. *medRxiv*, 2023.

[11] Alexandre Filiot, Paul Jacob, Alice Mac Kain, and Charlie Saillard. Phikon-v2, a large and public feature extractor for biomarker prediction, 2024. arXiv:2409.09173.

[12] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nat. Mach. Intell.*, 2(11):665–673, November 2020.

[13] Clément Grisi, Khrystyna Faryna, Nefise Uysal, Vittorio Agosti, Enrico Munari, Solène-Florence Kammerer-Jacquet, Paulo Guilherme de Oliveira Salles, Yuri Tolkach, Reinhard Büttner, Sofiya Semko, Maksym Pikul, Axel Heidenreich, Jeroen van der Laak, and Geert Litjens. Deep learning from routine histology improves risk stratification for biochemical recurrence in prostate cancer, 2026. arXiv:2603.14187.

[14] Frederick M Howard, James Dolezal, Sara Kochanny, Jefree Schulte, Heather Chen, Lara Heij, Dezheng Huo, Rita Nanda, Olufunmilayo I Olopade, Jakob N Kather, Nicole Cipriani, Robert L Grossman, and Alexander T Pearson. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nat. Commun.*, 12(1):4423, July 2021.

[15] Longlong Jing and Yingli Tian. Self-Supervised Visual Feature Learning With Deep Neural Networks: A Survey . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 43(11):4037–4058, November 2021.

[16] Saarthak Kapse, Mehmet Aygün, Elijah Cole, Emma Lundberg, Le Song, and Eric P. Xing. GenBio-PathFM: a state-of-the-art foundation model for histopathology. *bioRxiv*, 2026.

[17] 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, 2025. arXiv:2504.05186.

[18] Yousef Kotp, Vincent Quoc-Huy Trinh, Christopher Pal, and Mahdi S. Hosseini. MOOZY: A patient-first foundation model for computational pathology, 2026. arXiv:2603.27048.

[19] Jonah Kömen, Edwin D. de Jong, Julius Hense, Hannah Marienwald, Jonas Dippel, Philip Naumann, Eric Marcus, Lukas Ruff, Maximilian Alber, Jonas Teuwen, Frederick Klauschen, and Klaus-Robert Müller. Towards robust foundation models for digital pathology, 2025. arXiv:2507.17845.

[20] Jonah Kömen, Hannah Marienwald, Jonas Dippel, and Julius Hense. Do histopathological foundation models eliminate batch effects? a comparative study, 2024. arXiv:2411.05489.

[21] Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, Anil V. Parwani, Andrew Zhang, and Faisal Mahmood. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 2024.

[22] Jiabo Ma, Zhengrui Guo, Fengtao Zhou, Yihui Wang, Yingxue Xu, Jinbang Li, Fang Yan, Yu Cai, Zhengjie Zhu, Cheng Jin, Yi Lin, Xinrui Jiang, Chenglong Zhao, Danyi Li, Anjia Han, Zhenhui Li, Ronald Cheong Kin Chan, Jiguang Wang, Peng Fei, Kwang-Ting Cheng, Shaoting Zhang, Li Liang, and Hao Chen. A generalizable pathology foundation model using a unified knowledge distillation pretraining framework. *Nature Biomedical Engineering*, 10(3):545–564, 2026.

[23] Faisal Mahmood. A benchmarking crisis in biomedical machine learning. *Nature Medicine*, 31(4):1060, April 2025.

[24] Mahmood Lab. MahmoodLab/UNI2-h. Hugging Face model repository, https://huggingface.co/MahmoodLab/UNI2-h, 2025.

[25] Dmitry Nechaev, Alexey Pchelnikov, and Ekaterina Ivanova. Hibou: a family of foundational vision transformers for pathology, 2024. arXiv:2406.05074.

[26] Peter Neidlinger, Omar S M El Nahhas, Hannah Sophie Muti, Tim Lenz, Michael Hoffmeister, Hermann Brenner, Marko van Treeck, Rupert Langer, Bastian Dislich, Hans Michael Behrens, Christoph Röcken, Sebastian Foersch, Daniel Truhn, Antonio Marra, Oliver Lester Saldanha, and Jakob Nikolas Kather. Benchmarking foundation models as feature extractors for weakly supervised computational pathology. *Nat. Biomed. Eng.*, 10(6):1113–1123, June 2026.

[27] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision, 2024. arXiv:2304.07193.

[28] 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.

[29] Marin Scalbert, Charlie Saillard, Thomas Peeters, Liam Gonzalez, Dasha Valter, Felipe Llinares-López, Zelda E. Mariet, and Rodolphe Jenatton. Abstract LB174: H-optimus-1: a foundation model for computational histopathology. *Cancer Research*, 86(8_Supplement):LB174, 2026. Model weights: https://huggingface.co/bioptimus/H-optimus-1.

[30] George Shaikovski, Adam Casson, Kristen Severson, Eric Zimmermann, Yi Kan Wang, Jeremy D. Kunz, Juan A. Retamero, Gerard Oakley, David Klimstra, Christopher Kanan, Matthew Hanna, Michal Zelechowski, Julian Viret, Neil Tenenholtz, James Hall, Nicolo Fusi, Razik Yousfi, Peter Hamilton, William A. Moye, Eugene Vorontsov, Siqi Liu, and Thomas J. Fuchs. PRISM: a multi-modal generative foundation model for slide-level histopathology. *arXiv preprint arXiv:2405.10254*, 2024.

[31] Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, Ellen Yang, Philippe Mathieu, Alexander van Eck, Donghun Lee, Julian Viret, Eric Robert, Yi Kan Wang, Jeremy D Kunz, Matthew C H Lee, Jan H Bernhard, Ran A Godrich, Gerard Oakley, Ewan Millar, Matthew Hanna, Hannah Wen, Juan A Retamero, William A Moye, Razik Yousfi,

Christopher Kanan, David S Klimstra, Brandon Rothrock, Siqi Liu, and Thomas J Fuchs. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nat. Med.*, 30(10):2924–2935, October 2024.

[32] Jinxi Xiang, Xiyue Wang, Xiaoming Zhang, Yinghua Xi, Feyisope Eweje, Yijiang Chen, Yuchen Li, Colin Bergstrom, Matthew Gopaulchan, Ted Kim, Kun-Hsing Yu, Sierra Willens, Francesca Maria Olguin, Jeffrey J. Nirschl, Joel Neal, Maximilian Diehn, Sen Yang, and Ruijiang Li. A vision–language foundation model for precision oncology. *Nature*, 638(8051):769–778, 2025.

[33] Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, Yanbo Xu, Mu Wei, Wenhui Wang, Shuming Ma, Furu Wei, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Jaylen Rosemon, Tucker Bower, Soohee Lee, Roshanthi Weerasinghe, Bill J Wright, Ari Robicsek, Brian Piening, Carlo Bifulco, Sheng Wang, and Hoifung Poon. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, June 2024.

[34] Yingxue Xu, Yihui Wang, Fengtao Zhou, Jiabo Ma, Cheng Jin, Shu Yang, Jinbang Li, Zhengyu Zhang, Chenglong Zhao, Huajun Zhou, Zhenhui Li, Huangjing Lin, Xin Wang, Jiguang Wang, Anjia Han, Ronald Cheong Kin Chan, Li Liang, Xiuming Zhang, and Hao Chen. A multimodal knowledge-enhanced whole-slide pathology foundation model. *Nature Communications*, 16(1):11406, 2025.

[35] Eric Zimmermann, Eugene Vorontsov, Julian Viret, Adam Casson, Michal Zelechowski, George Shaikovski, Neil Tenenholtz, James Hall, David Klimstra, Razik Yousfi, Thomas Fuchs, Nicolo Fusi, Siqi Liu, and Kristen Severson. Virchow2: Scaling self-supervised mixed magnification models in pathology, 2024. arXiv:2408.00738.

A Supplementary material

A.1 Example samples

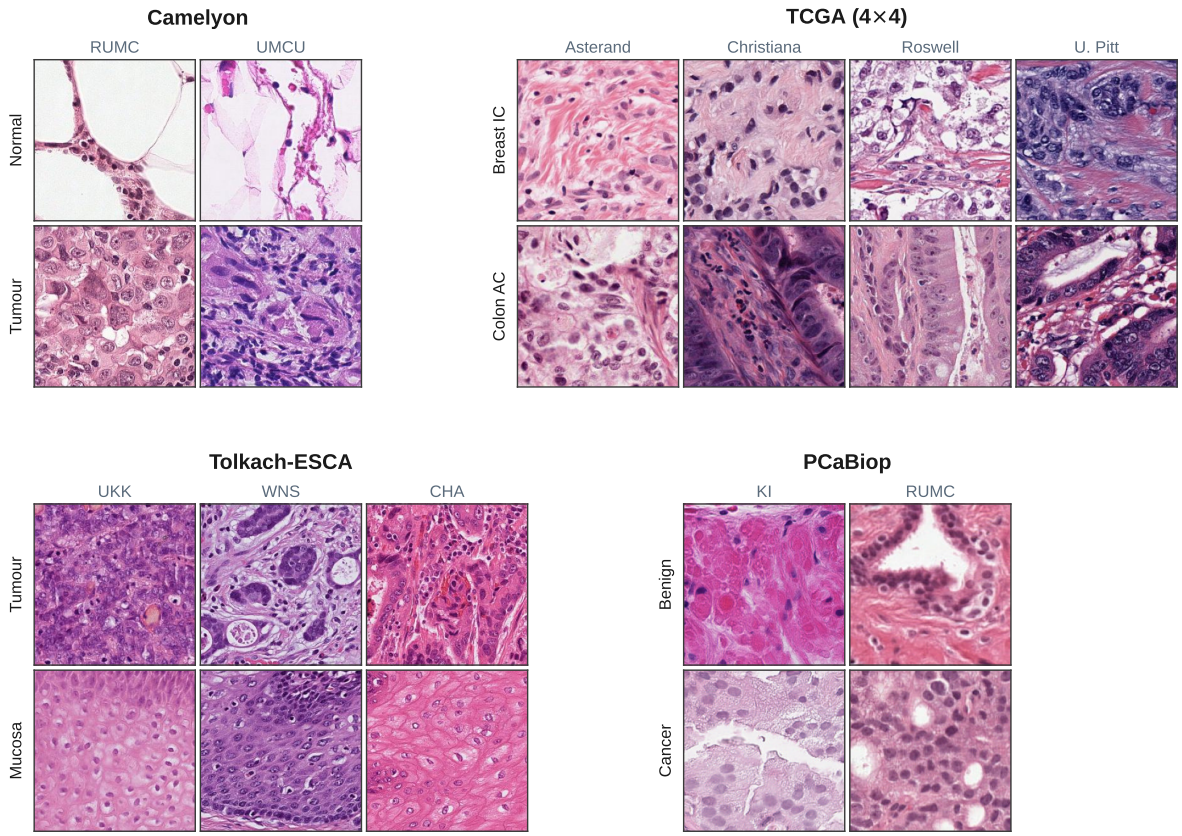


Figure 7: **Biology-matched example samples across acquisition centres.** Each benchmark block shows two biological classes (rows) across all acquisition centres (columns), so within a row biology is fixed and only the centre changes. Because PCaBiop is a slide-level benchmark, its block shows representative tiles rather than whole slides, which make the visual differences between its centres easier to see. The resulting staining, tissue preparation and scanning differences are well visible: Camelyon's UMCU tiles are markedly more purple than RUMC's, and the three Tolkach-ESCA cohorts differ in overall hue, exactly the non-biological variation a site-invariant representation must not treat as signal.

A.2 Sensitivity of $\texttt{LTM}_\alpha$ to the tail fraction α

The tail-severity summary $\texttt{LTM}_\alpha$ requires a tail fraction α, fixed at $\alpha = 0.10$ in the main text. Here we show that model comparison by $\texttt{LTM}_\alpha$ is insensitive to this choice over a plausible range. For each benchmark we recompute $\texttt{LTM}_\alpha$ per model at $\alpha \in \{0.05, 0.10, 0.20\}$ from the per-sample $\texttt{CRoMa}(m{=}5)$ values, rank the models, and measure the Spearman rank correlation between the rankings induced by different α. Rankings are stable between adjacent tail fractions on every benchmark (Table 7): the 0.05-vs-0.10 and 0.10-vs-0.20 correlations are ≥ 0.89 throughout, and the 0.05-vs-0.10 correlation exceeds 0.94 on four of the five benchmarks. As expected, the widest

comparison (0.05 vs 0.20) is looser because the 5% and 20% tails probe increasingly different subpopulations. The default $\alpha = 0.10$ therefore sits in a stable interior region, and the tail-analysis conclusions do not hinge on it.

Benchmark	$\rho(0.05, 0.10)$	$\rho(0.10, 0.20)$	$\rho(0.05, 0.20)$
Camelyon	0.98	0.99	0.94
TCGA-2x2	0.98	0.94	0.89
TCGA-4x4	0.96	0.90	0.77
Tolkach-ESCA	0.90	0.89	0.69
PCaBiop	1.00	1.00	1.00

Table 7: **Rank correlations across lower-tail fractions.** Entries are Spearman correlations between per-model $\texttt{LTM}_\alpha$ rankings at $\alpha = 0.05$, 0.10 and 0.20, computed from `CRoMa`(m=5) within each benchmark. Each tile benchmark ranks 20 encoders. PCaBiop ranks 4, so its correlations rest on a far smaller panel.

A.3 Robustness of `CRoMa` model comparison to the averaging radius m

`CRoMa` averages the distances to the m nearest neighbours of each type, fixed at $m = 5$ in the main text. Here we show that model comparison is insensitive to this choice over a wide range of m. For the three tile benchmarks (Camelyon, TCGA-4x4, and Tolkach-ESCA) we take the pooled `CRoMa` per model at every integer $m \in [1, 20]$ and track both the ranking and the number of confounder-dominant models (`CRoMa` < 0). The comparison is essentially invariant across the sweep (Table 8). Model rankings are highly concordant for every benchmark (Spearman $\rho \geq 0.98$ between the extremes $m = 1$ and $m = 20$), and the number of confounder-dominant models is constant across the full sweep. Increasing m therefore affects only margin magnitudes – and the lower-tail summaries derived from them – without altering model order or the sign of the dominant signal. Thus, $m = 5$ represents a practical operating point rather than a tuned parameter: averaging over five neighbours limits the contribution of any single neighbour to one-fifth of either type-specific mean.

Benchmark	n	$\rho(m{=}1, m{=}5)$	$\rho(m{=}5, m{=}20)$	$\rho(m{=}1, m{=}20)$
Camelyon	20	0.992	0.989	0.982
TCGA (4×4)	20	0.992	0.992	0.983
Tolkach-ESCA	20	0.994	0.992	0.988

Table 8: **Robustness of `CRoMa` model comparison to the averaging radius m.** For each of the three tile benchmarks, we report Spearman correlations between the median `CRoMa` model rankings at the headline radius $m = 5$ and each extreme ($m = 1$, $m = 20$), and between the two extremes.

A.4 Why the selected neighbourhood size undercuts fixed-k coverage

Fixed-k metrics derive their neighbourhood size from a biological k-NN criterion: k is chosen to maximise biological k-NN classification accuracy (Methods). A biological k-NN classifier votes on the biology label of *every* neighbour in the window. Since same-biology, same-centre samples are naturally closer than other neighbour types, the classifier's accuracy is dominated by the dense `SS` pocket surrounding each anchor. On Camelyon, `SS` neighbours comprise 86–93% of the selected

window. Maximising that accuracy therefore drives the selected size small (median $k = 11$ on Camelyon, per-model optima 7–67), toward the very SS neighbours the robustness metrics discard as uninformative and rarely extending far enough to capture the typed SO/OS neighbours they score. The first SO or OS neighbour occurs only at a median rank of approximately 149, far beyond the operated $k = 11$. The coverage gap reported in the main text follows directly: a window selected by an SS-dominated classification objective is, almost by construction, too small to include the typed neighbours required by RI. Most anchors therefore contain no typed evidence and are silently excluded from the pooled score. CRoMa avoids this conflict by imposing no fixed neighbourhood at all, reading each typed neighbour at whatever rank it occurs (Section 3.2).

A.5 A geometric interpretation of the CRoMa margin

Because $\texttt{CRoMa}_i$ depends only on the ratio $d_m^{\texttt{OS}}/d_m^{\texttt{SO}}$, it has a simple reading in the plane of typed mean distances $(d_m^{\texttt{SO}}, d_m^{\texttt{OS}})$ (Figure 8). Writing this pair in polar coordinates, with

$$\theta_i = \operatorname{atan2}\big(d_m^{\texttt{OS}}, d_m^{\texttt{SO}}\big)\,,$$

gives

$$\texttt{CRoMa}_i(m) = \tan\Big(\theta_i - \frac{\pi}{4}\Big)\,.$$

Thus, $\texttt{CRoMa}_i$ measures the signed angular deviation from the 45° diagonal. Equal typed distances lie on this diagonal, $d_m^{\texttt{SO}} = d_m^{\texttt{OS}}$, and yield $\texttt{CRoMa}_i(m) = 0$. Points above the diagonal, where the distractor is farther away, are robust ($\texttt{CRoMa}_i(m) > 0$), and points below it are fragile ($\texttt{CRoMa}_i(m) < 0$). The score is independent of radial distance: moving a point along a ray from the origin scales both typed distances by a common factor $\lambda > 0$, which cancels in the ratio $d_m^{\texttt{OS}}/d_m^{\texttt{SO}}$ and leaves θ_i—and hence $\texttt{CRoMa}_i$—unchanged, so all points on a ray share the same margin. This cancellation is precisely the scale-free property noted in 2.4: only the *direction* of the typed-distance pair carries signal, not its length.

This scale invariance is what lets CRoMa margins be compared across settings that hold one factor fixed and vary the other. The absolute value of $d_m^{\texttt{SO}}$ and $d_m^{\texttt{OS}}$ is a nuisance quantity: it reflects the encoder (embedding norm, output temperature, feature dimension) and the benchmark (biological classes, cohort composition) rather than robustness itself. With the benchmark fixed and the model varied, an encoder that merely spreads all of its distances wider is not credited as more robust, because the shared factor λ cancels and only the relative separation of cross- and same-confounder neighbours—the angle θ_i—remains. With the model fixed and the benchmark varied, datasets that sit at different typical distance magnitudes are placed on one common scale, so a model's margin on Camelyon reads on the same footing as its margin on TCGA-4x4.

The nuisance is large in practice, not hypothetical. Measuring the plane directly on PCaBiop (Figure 9) places the 4 slide-level encoders at median typed-distance radii spanning a factor of 80, so their point clouds sit at four widely separated positions along the diagonal. Their robustness is unrelated to that position. PRISM and TITAN make the point sharply: they occupy nearly the same radius yet fall on opposite sides of the diagonal. Reading robustness off the raw typed distances would therefore rank encoders largely by embedding geometry. Our ratio removes that dependence by construction.

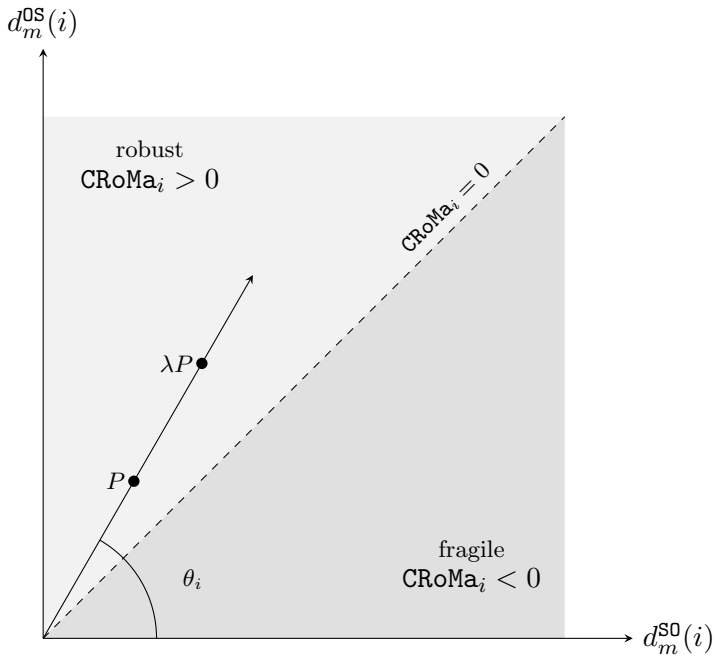


Figure 8: **CRoMa is the angular offset from the diagonal.** In the $(d_m^{\mathtt{SO}}, d_m^{\mathtt{OS}})$ plane, $\mathtt{CRoMa}_i = \tan(\theta_i - \pi/4)$: points above the diagonal have more distant OS distractors and positive margins, points below have negative margins. P and λP lie on one ray, so they share θ_i and therefore the same margin.

A.6 Metric rank agreement within each benchmark

Benchmark	RI vs MaRI	CRoMa vs RI	CRoMa vs MaRI
Camelyon	0.99	0.95	0.94
TCGA 4x4	0.99	0.95	0.95
Tolkach-ESCA	0.98	0.93	0.92

Table 9: **Pairwise rank correlations among RI, MaRI and CRoMa.** Spearman correlations across the 20 pathology encoders shared by the three tile benchmarks. The three metrics induce nearly identical model orderings on every tile benchmark: RI and MaRI are almost interchangeable as rankers, and CRoMa tracks both wherever the fixed-k scores are defined, despite scoring a different and larger population of anchors. A shared ranking is not interchangeability, however: encoders with matched MaRI can separate sharply under CRoMa (Section 3.2), and only the per-sample construction exposes the confounder-dominated tail (Section 3.3).

A.7 Pretraining provenance

Midnight-12k was pretrained exclusively on TCGA, so its wide lead on TCGA-4x4—CRoMa=0.40 versus 0.16 for the runner-up, GenBio-PathFM—could partly reflect pretraining overlap. To test this, we re-scored a larger Tolkach-ESCA cohort that adds the held-out TCGA cases and measured the resulting robustness gains (Supplementary Table 10). Midnight-12k shows the largest TCGA-cohort boost of any encoder (2.5× in typed-distance odds), but exposure alone does not explain it: every encoder gains on this cohort, including the natural-image control (1.21×), and the 9

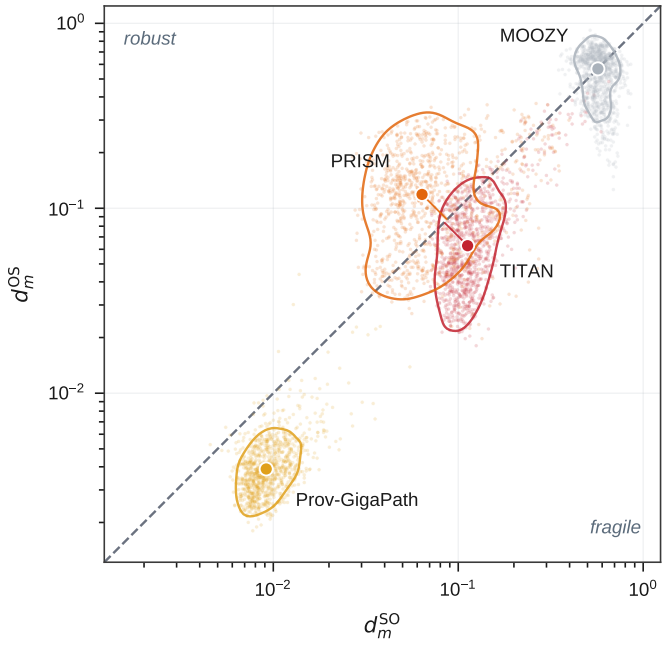


Figure 9: **The typed-distance plane measured on PCaBiop.** One point per slide for the 4 slide-level encoders ($m = 5$). Outlines enclose 75% of an encoder's slides and filled markers are per-encoder medians. Axes are logarithmic because median typed-distance radii span a factor of 80, from 0.010 (`Prov-GigaPath`) to 0.80 (`MOOZY`). The log map turns a common rescaling into a shift along the diagonal, so position along the diagonal is scale and offset from it is robustness: the angular offset $\theta_i - \pi/4$ of Figure 8 becomes a perpendicular displacement from the diagonal of length $|\log r|/\sqrt{2}$, where $r = d_m^{\mathtt{OS}}/d_m^{\mathtt{SO}}$, the side of the diagonal carrying its sign. `PRISM` and `TITAN` illustrates this best: both models sit at nearly the same radius (0.14 versus 0.13) but on opposite sides of the diagonal (`CRoMa` $= 0.26$ versus -0.30). The plane also separates failure modes the sign alone conflates: `Prov-GigaPath` is fragile with all typed distances collapsed toward zero, whereas `TITAN` is fragile at an ordinary distance scale.

TCGA-exposed encoders are barely separated from the rest (median 1.36× versus 1.26×). `Phikon`, also pretrained exclusively on public TCGA, gains only 1.35×. `Midnight-12k` still leads once the TCGA cases are removed. Pretraining overlap therefore amplifies, rather than creates, the observed advantage.

Model	Tolkach-ESCA	TCGA extension	Boost
Midnight-12k†	0.60	0.82	2.50×
H0-mini†	0.38	0.53	1.47×
Virchow	0.37	0.52	1.45×
Virchow2	0.37	0.51	1.44×
Hibou-L	0.12	0.28	1.41×
UNI2-h	0.23	0.38	1.39×
GenBio-PathFM†	0.41	0.53	1.37×
GPFM†	0.24	0.38	1.37×
Prost40M†	0.14	0.28	1.36×
Phikon†	0.17	0.31	1.35×
H-optimus-0	0.24	0.37	1.32×
MUSK†	0.30	0.42	1.32×
Phikon-v2†	0.13	0.26	1.31×
mSTAR†	0.20	0.32	1.28×
Prov-GigaPath	0.13	0.24	1.26×
H-optimus-1	0.27	0.38	1.26×
CONCH	0.44	0.52	1.24×
CONCHv1.5	0.41	0.49	1.22×
Hibou-B	0.14	0.23	1.22×
UNI	0.18	0.25	1.17×
DINOv2-B	0.17	0.27	1.21×

Table 10: **Median `CRoMa` on Tolkach-ESCA and on its TCGA extension.** Median per-sample `CRoMa`(m=5) for the 20 tile-level pathology encoders, computed separately over the three original Tolkach-ESCA cohorts and over the held-out TCGA cases. The *boost* measures how much further the nearest same-confounder biological distractors sit (relative to the nearest cross-confounder biological matches) on the TCGA cases than on the original cohorts. It is the between-cohort ratio of $r = d_m^{\mathtt{OS}}/d_m^{\mathtt{SO}}$, the typed-distance ratio on which `CRoMa` is defined (Section 2.4). † marks the 9 TCGA-exposed encoders (Table 2). The natural-image control `DINOv2-B` is shown separately.

A.8 CRoMa distributions

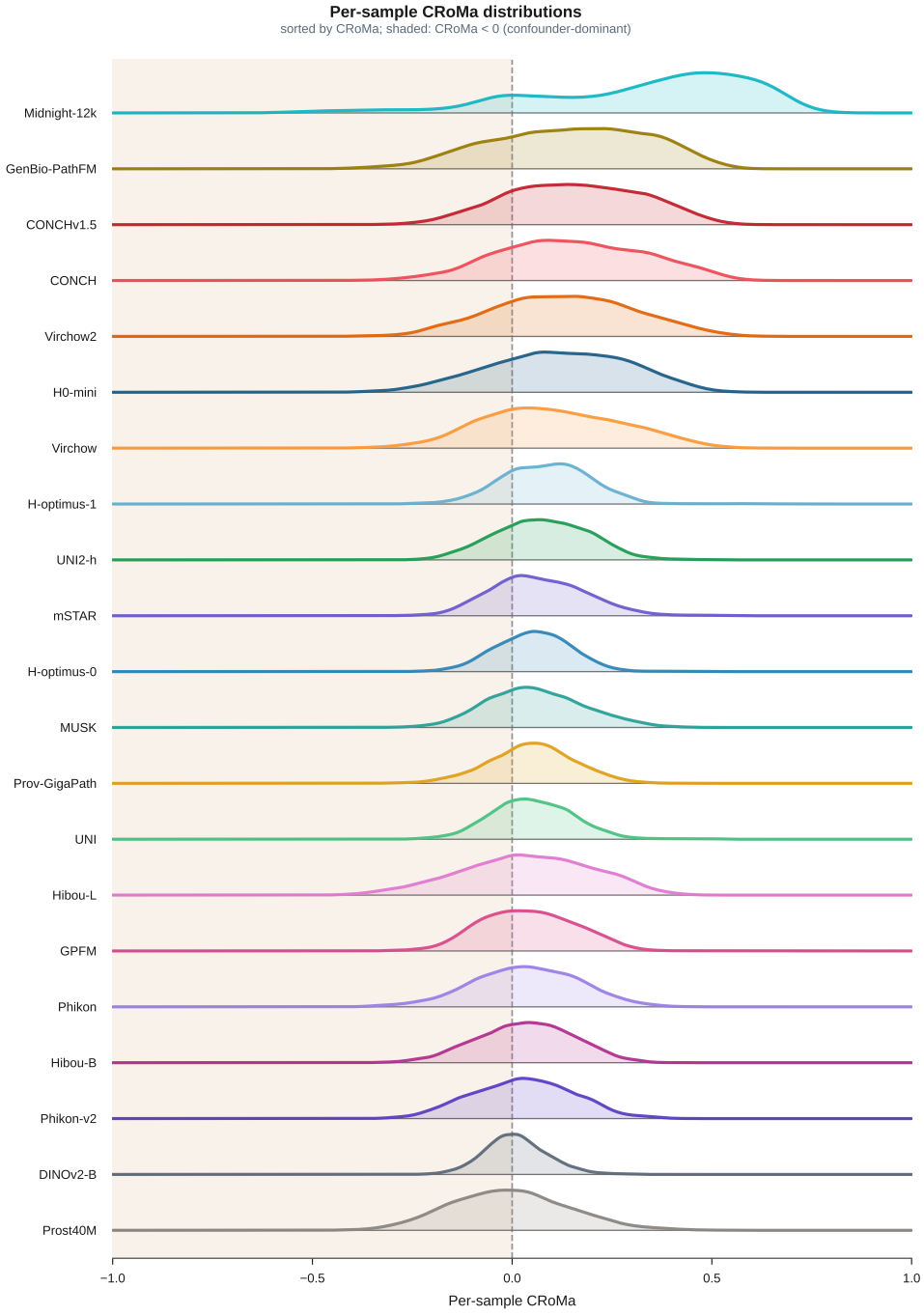


Figure 10: **Per-sample CRoMa distributions on TCGA-4x4.** Ridgeline distributions for 20 pathology encoders and the natural-image control (DINOv2-B), ordered by pooled median. The dashed line denotes CRoMa $= 0$; shading denotes CRoMa < 0. Per-model $F(0)$ and LTM_{10} are reported in Table 4.

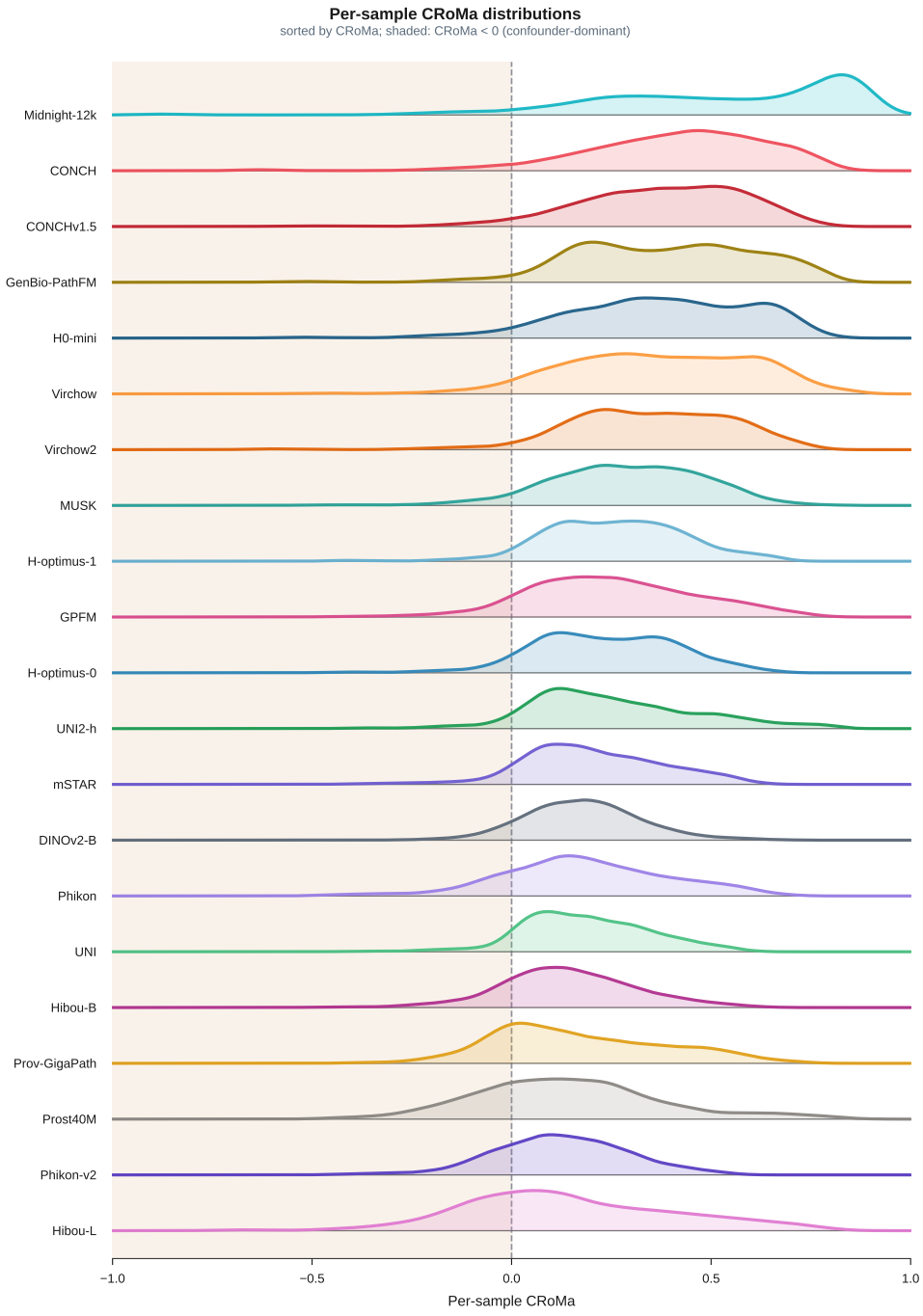


Figure 11: **Per-sample `CRoMa` distributions on Tolkach-ESCA.** Ridgeline distributions for 20 pathology encoders and the natural-image control (`DINOv2-B`), ordered by pooled median. The dashed line denotes $\texttt{CRoMa} = 0$; shading denotes $\texttt{CRoMa} < 0$. Per-model $F(0)$ and $\texttt{LTM}_{10}$ are reported in Table 5.

A.9 Downstream shortcut susceptibility

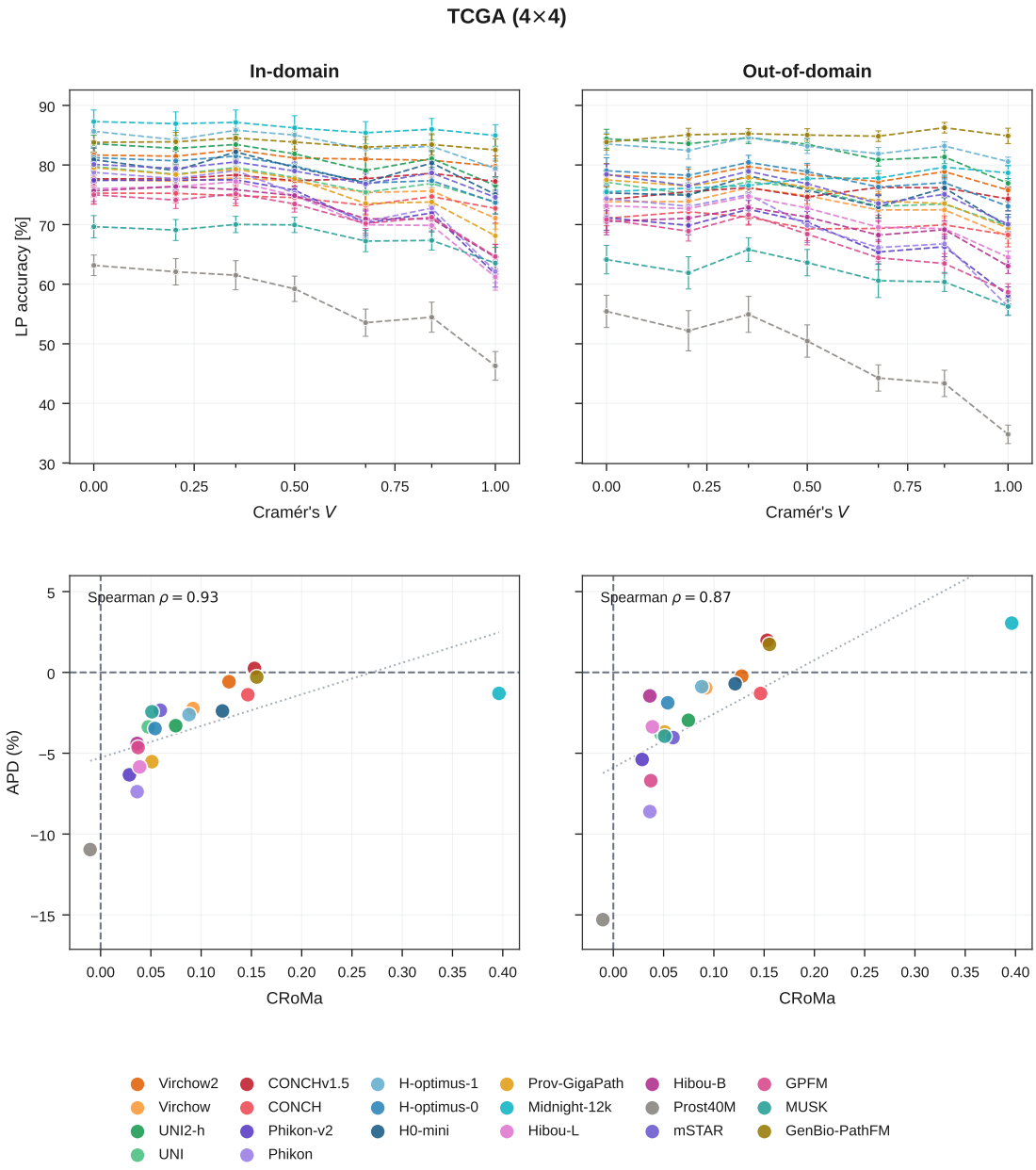


Figure 12: **CRoMa and APD on TCGA-4x4.** Columns show in-domain evaluation (left; test centres held fixed) and out-of-domain evaluation (right; unseen centres). **Top row,** balanced-test linear-probe accuracy as the centre–biology correlation in the training set increases from V=0 to V=1; dashed curves denote the 20 tile-level encoders and error bars are 95% t-intervals over seeds. **Bottom row,** pooled CRoMa(m=5) versus APD, one point per encoder. Dashed lines mark CRoMa = 0 and APD = 0. Corresponding RI and MaRI correlations are reported in Table 11.

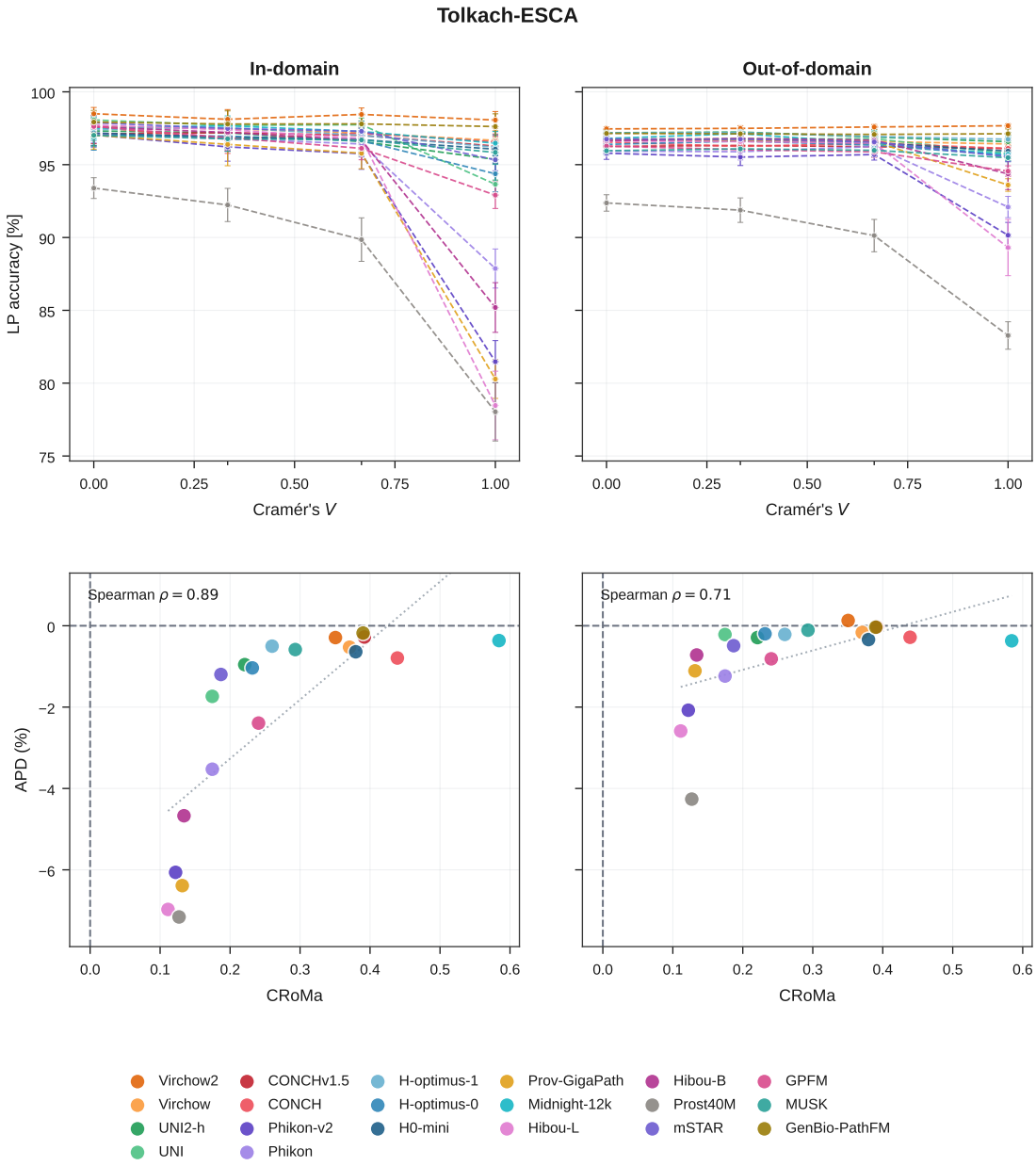


Figure 13: `CRoMa` **against** `APD` **on Tolkach-ESCA.** Columns show in-domain evaluation (left; test centres held fixed) and out-of-domain evaluation (right; unseen centres). **Top row,** balanced-test linear-probe accuracy as the centre–biology correlation in the training set increases from V=0 to V=1; dashed curves denote the 20 tile-level encoders and error bars are 95% t-intervals over seeds. **Bottom row,** pooled `CRoMa`(m=5) versus `APD`, one point per encoder. Dashed lines mark `CRoMa` = 0 and `APD` = 0. Corresponding `RI` and `MaRI` correlations are reported in Table 11.

	Metric	Camelyon	TCGA-4x4	Tolkach-ESCA	*pooled*
$\texttt{APD}_{\text{ID}}$	`CRoMa`	0.88	0.93	0.89	0.84
	`RI`	0.87	0.87	0.97	0.89
	`MaRI`	0.86	0.89	0.95	0.89
$\texttt{APD}_{\text{OOD}}$	`CRoMa`	0.71	0.87	0.71	0.78
	`RI`	0.73	0.86	0.85	0.76
	`MaRI`	0.70	0.85	0.84	0.74

Table 11: **Correlation between representation robustness metrics and downstream shortcut susceptibility**. Spearman ρ across the 20 pathology encoders between the three robustness metrics (`RI`, `MarRI` and `CRoMa`) and downstream shortcut susceptibility (`APD`). Correlations are reported for each benchmark, as well as computed jointly over all 60 model–benchmark pairs (*pooled* column). Positive values indicate that larger robustness scores are associated with smaller performance drops. $\texttt{APD}_{\text{ID}}$ holds test centres fixed. $\texttt{APD}_{\text{OOD}}$ uses unseen centres. All three metrics track `APD` comparably. Correlations weaken out of domain for all three, where `APD` additionally reflects transfer to unseen confounders rather than shortcut reliance alone.

A.10 Median–tail Pareto frontiers

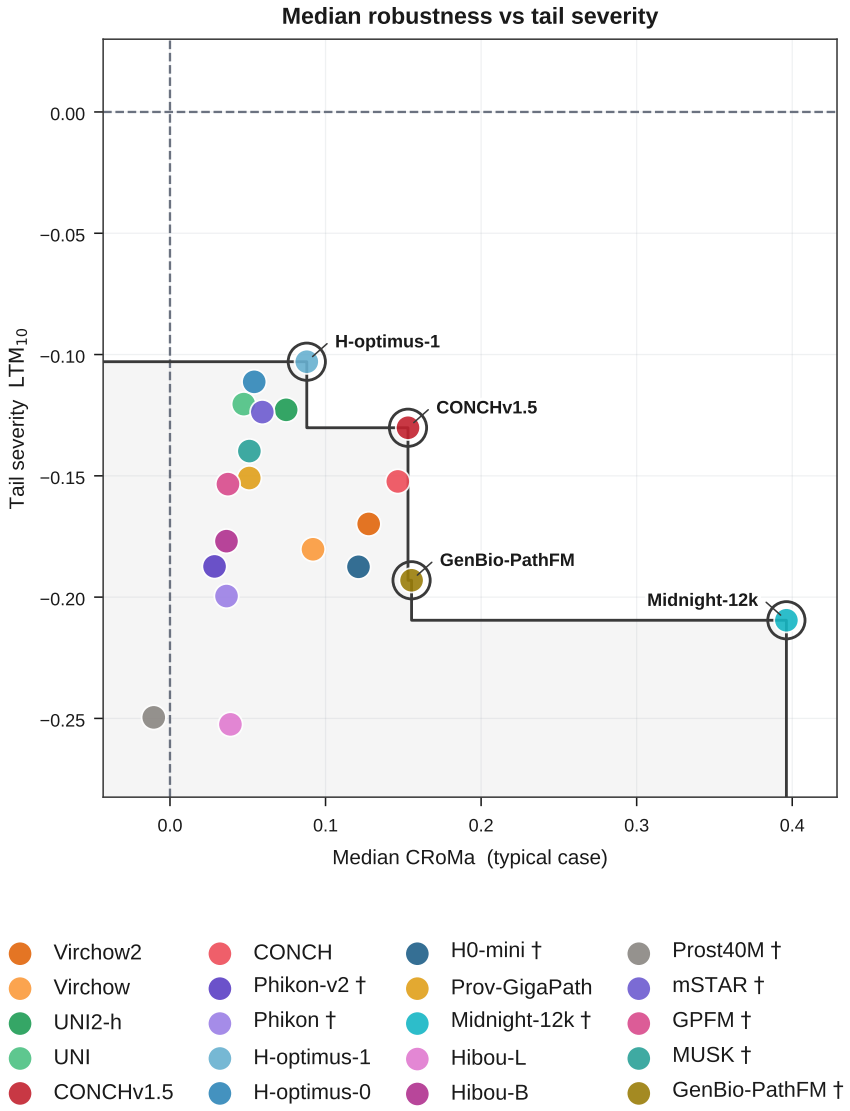


Figure 14: **Median and lower-tail `CRoMa` on TCGA-4x4.** Pooled median `CRoMa` against worst-decile mean $\texttt{LTM}_{10}$ for the 20 pathology encoders. Higher values are preferable on both axes. Ringed points form the upper-right Pareto frontier, while shaded points are dominated on both axes. Per-model $F(0)$ and $\texttt{LTM}_{10}$ are reported in Table 4. † marks the 9 TCGA-exposed encoders.

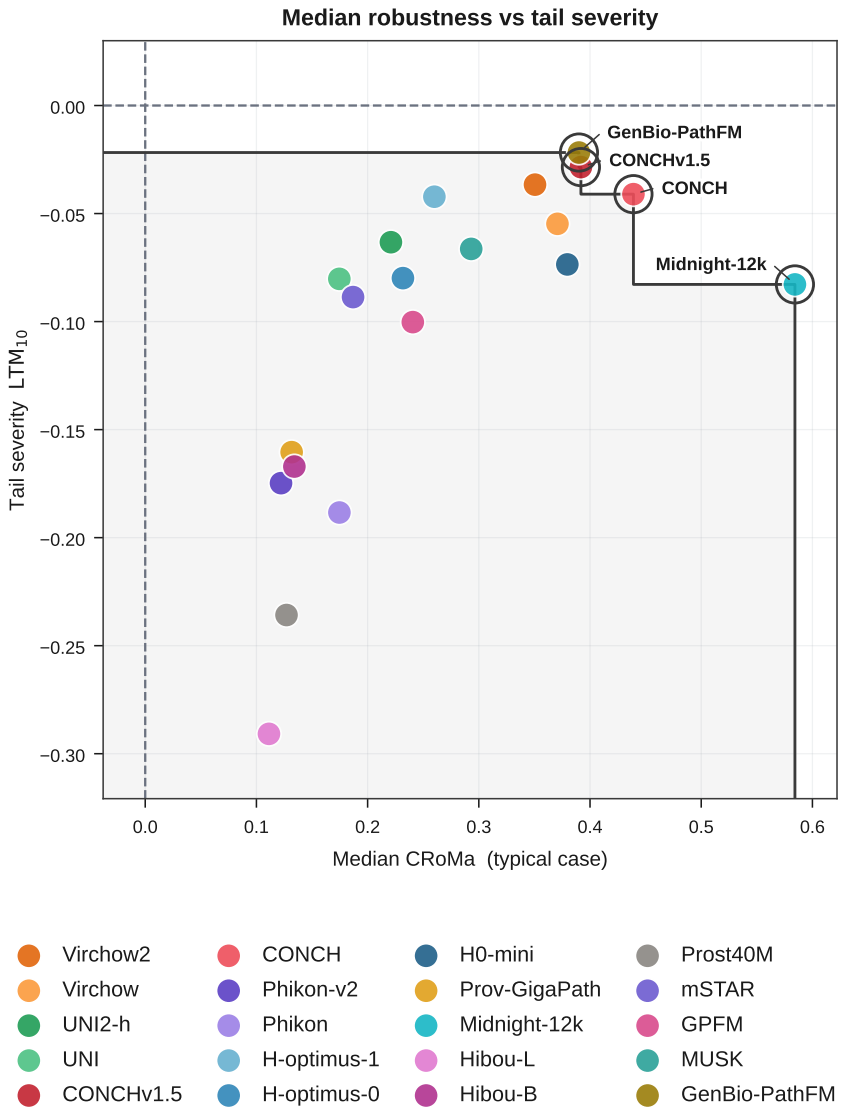


Figure 15: **Median and lower-tail `CRoMa` on Tolkach-ESCA.** Pooled median `CRoMa` against worst-decile mean $\texttt{LTM}_{10}$ for the 20 pathology encoders. Higher values are preferable on both axes. Ringed points form the upper-right Pareto frontier, while shaded points are dominated on both axes. Per-model $F(0)$ and $\texttt{LTM}_{10}$ are reported in Table 5.

		median rank	
Model	CRoMa	SO	OS
Virchow2	0.20	111	476
CONCH	0.20	97	312
GenBio-PathFM	0.19	114	694
CONCHv1.5	0.19	55	311
H0-mini	0.17	134	393
Virchow	0.16	194	450
Midnight-12k	0.11	241	338
H-optimus-1	0.08	276	550
H-optimus-0	0.05	261	367
UNI2-h	0.04	379	570
MUSK	0.04	195	197
mSTAR	0.02	329	338
Prov-GigaPath	0.01	445	376
UNI	-0.03	933	414
Hibou-B	-0.09	1211	304
GPFM	-0.10	785	193
Phikon	-0.20	2216	313
Phikon-v2	-0.21	2909	269
Prost40M	-0.32	1817	124
Hibou-L	-0.44	9089	327
DINOv2-B	0.05	32	50

Table 12: **Typed-neighbour ranks on Camelyon.** Models are ordered by pooled CRoMa(m=5). Median SO and OS ranks summarize, across models, the rank of the nearest cross-confounder biological match and same-confounder biological distractor, respectively. Pooled across the 20 encoders, the first SO or OS neighbour occurs at a median rank of $\approx$ 149 among 20,400 candidates, sitting far beyond the k=11 operating point selected by the biological k-NN criterion on Camelyon. This neighbourhood size is therefore too narrow for the fixed-k metrics RI and MaRI to capture the typed SO/OS contrast they need. DINOv2-B, the natural-image control, is reported separately for reference.

A.11 TCGA: the paired 2x2 configuration

For completenes, we additionally report PathoROB's paired TCGA-2x2 protocol here (Table 13; composition in Figure 16): 94 balanced quartets, each two biological classes $\times$ two centres, with neighbour-type statistics computed within each quartet and pooled at the level of SO/OS counts before scoring. Results on this paired configuration align with the observations made on the TCGA-4x4 benchmark: biology-dominant pooled scores and a confounder-dominant lower tail for every model, with Midnight-12k leading (CRoMa 0.42 in both).

A.12 Robustness of slide-level representations

Unlike the tile-level panels, the four slide-level encoders were evaluated at their model-specific, biologically selected $k^\star$. With only four models, the median (the lower of the two central optima) is highly sensitive to panel composition and collapses to $k = 3$ on PCaBiop, reducing mean RI/MaRI support across the four encoders from 36.9% at per-model $k^\star$ to 27.0%. We therefore treat the slide-level panel as an explicit exception to the shared median-k protocol. This preserves each encoder's biologically selected neighbourhood scale, although comparisons of RI, MaRI and support remain

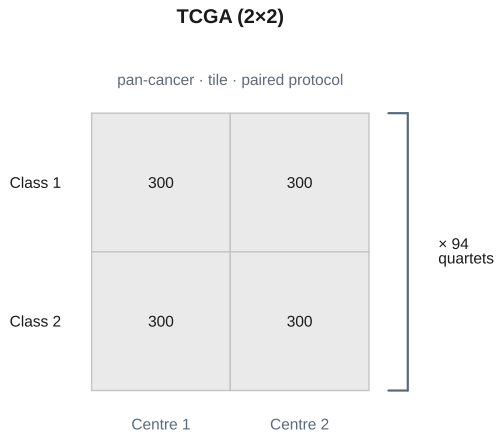


Figure 16: **The paired TCGA protocol comprises 94 balanced 2×2 quartets.** The grid shows one generic quartet: two biological classes by two medical centres, with 300 tile occurrences per cell. Class and centre identities vary across quartets, spanning 21 cancer types and 26 centres. Typed-neighbour counts are pooled across quartets before scoring, for 112,800 tile occurrences in total. The dataset-wide TCGA-4x4 design is shown in Figure 2.

conditional on model-specific k. `CRoMa`, $F(0)$ and $\texttt{LTM}_{10}$ are k-independent and are unaffected by this choice.

Binary cancer detection. On PCaBiop, the contributing centre was near-perfectly decodable for all four encoders (confounder k-NN balanced accuracy, 0.992–1), whereas biological balanced accuracy ranged from 0.716 to 0.971 (Supplementary Table 14). Model-level median `CRoMa` values spanned $[-0.41, 0.26]$, with only `PRISM` exhibiting biology-dominant neighbourhood geometry (0.26). `MOOZY` [18] attained the highest biological balanced accuracy (0.971), but its median margin remained slightly negative (-0.02), illustrating that high biological discriminability does not imply robustness to confounder. The per-sample distributions provided complementary information (Supplementary Figure 17). `PRISM` and `MOOZY` differed in median margin (0.26 versus -0.02) and failure prevalence (28.8% versus 53.5%), but had similar worst-decile severity ($\texttt{LTM}_{10} = -0.39$ versus -0.41). Whereas the Camelyon analysis distinguished encoders with similar medians but different lower tails, this comparison holds lower-tail severity approximately constant while median robustness and failure prevalence differ, further illustrating that the three summaries are non-redundant.

ISUP grading. To assess a more granular biological label, we curated PCaBiop-ISUP, which comprises 3,000 prostate biopsies. Each biopsy comes with its ISUP grade group: grade 0 denotes benign samples and grades 1–5 denote increasingly aggressive prostate cancer. Whole slides are still sampled from PANDA, so the confounder remains the medical center that provided the biopsy (KI vs RUMC). Each ISUP×center cell has exactly 250 biopsies (Supplementary Figure 18). Cell balance alone does not, however, balance the typed candidate pools in the full 6×2 cohort. For any anchor, the `SO` pool contains 250 slides of the same grade from the other center, whereas the `OS` pool contains 1,250 slides from the five other grades at the same center. This asymmetry would structurally favour `OS` evidence in both fixed-k and nearest-neighbour scores. We therefore decom-

Model	bio bacc	conf bacc	RI	MaRI	Δ	CRoMa	$F(0)$	LTM_{10}	support
Midnight-12k†	0.934	0.739	**0.858**	**0.890**	+0.032	**0.42**	**0.124**	-0.20	93.3%
GenBio-PathFM†	0.943	0.763	0.840	0.859	+0.020	0.27	0.148	-0.15	92.4%
CONCHv1.5	0.921	0.705	0.832	0.855	+0.023	0.26	0.144	-0.13	98.9%
CONCH	0.912	0.679	0.824	0.849	+0.025	0.25	0.154	-0.15	**99.1%**
Virchow2	0.928	0.724	0.822	0.841	+0.019	0.24	0.157	-0.15	97.6%
H0-mini†	0.925	0.743	0.792	0.810	+0.018	0.19	0.194	-0.17	95.6%
Virchow	0.909	0.752	0.761	0.775	+0.014	0.17	0.235	-0.26	95.1%
H-optimus-1	**0.949**	0.762	0.844	0.869	+0.026	0.17	0.144	**-0.09**	95.1%
UNI2-h	0.943	0.816	0.796	0.821	+0.025	0.13	0.182	-0.11	91.8%
H-optimus-0	0.938	0.778	0.794	0.819	+0.026	0.11	0.193	-0.11	94.3%
mSTAR†	0.929	0.799	0.765	0.778	+0.013	0.11	0.213	-0.13	95.4%
MUSK†	0.887	0.759	0.718	0.726	+0.008	0.10	0.248	-0.17	98.2%
Prov-GigaPath	0.925	0.821	0.728	0.753	+0.025	0.10	0.249	-0.14	92.6%
UNI	0.928	0.833	0.729	0.744	+0.015	0.09	0.243	-0.12	92.4%
Phikon†	0.913	0.884	0.600	0.604	+0.004	0.06	0.360	-0.22	81.8%
GPFM†	0.899	0.844	0.645	0.643	−0.002	0.06	0.318	-0.18	94.9%
Hibou-B	0.909	0.871	0.607	0.605	−0.002	0.06	0.361	-0.22	86.1%
Hibou-L	0.904	0.889	0.559	0.528	−0.031	0.05	0.422	-0.38	78.5%
Phikon-v2†	0.918	0.890	0.587	0.588	+0.001	0.05	0.382	-0.22	81.6%
Prost40M†	0.839	0.850	0.508	0.500	−0.008	0.01	0.463	-0.34	92.5%
DINOv2-B	0.802	0.706	0.612	0.604	−0.008	0.04	0.350	-0.16	99.9%

Table 13: **Representation robustness on TCGA-2x2.** Pooled results for 20 tile-level pathology foundation models, ordered by CRoMa (m=5), with the natural-image control DINOv2-B shown separately beneath. All models are evaluated at the shared operating point k=61, the dataset median of the per-model biological $k^\star$. Columns: biological and confounder k-NN balanced accuracy (bio bacc and conf bacc; confounder: medical centre); pooled RI and MaRI; Δ=MaRI − RI; median CRoMa; $F(0)$, the fraction with CRoMa < 0; LTM_{10}, the mean of the lowest decile; and support, the fraction of samples effectively contributing to RI/MaRI. Bold denotes the best value in each score column (conf bacc and Δ are diagnostics). † marks the 9 TCGA-exposed encoders (Table 2).

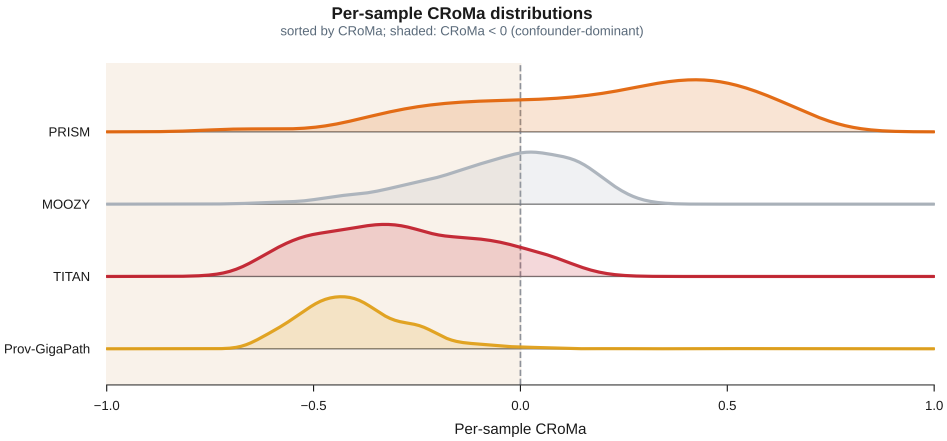


Figure 17: **CRoMa distributions on PCaBiop.** Ridgeline distributions for the 4 slide-level encoders on PCaBiop, ordered by pooled median. The dashed line denotes CRoMa $= 0$; shading denotes CRoMa < 0. Per-model $F(0)$ and LTM_{10} are reported in Table 14.

posed the six grades into all $\binom{6}{2} = 15$ grade-center pairs. Crossing each pair with the two center yielded a balanced quartets of 1,000 slides, with 250 `SO` and 250 `OS` candidates available to every anchor. `RI` and `MaRI` pooled typed-neighbour evidence across occurrences. To weight each grade contrast equally, the headline `CRoMa` was the median of the 15 pair-specific medians. $F(0)$ and $\texttt{LTM}_{10}$ retained their occurrence-level interpretations and were computed over all occurrence-level margins, with $m = 5$ throughout. PCaBiop is a matched 1,000-slide subset of the broader 3,000-slide PCaBiop-ISUP cohort. It contains both complete grade-0 cells as the benign class and, for the cancer class, 250 slides total per provider sampled across grades 1–5. The analyses differ in label resolution, sample composition and evaluation design. They therefore provide related, but not independent, evidence.

On PCaBiop-ISUP, the medical center remained near-perfectly decodable across the four encoders (confounder balanced accuracy, 0.990–1), and every encoder had a negative median `CRoMa` (Table 14). `MOOZY` combined the highest biological balanced accuracy (0.934) with the least negative median margin (−0.09). `PRISM`, however, had a slightly lower failure prevalence than `MOOZY` (65.1% versus 67.5%), again distinguishing median robustness from the prevalence of confounder-dominant occurrences. Because PCaBiop-ISUP also differs from binary PCaBiop in sample composition and evaluation design, the shift towards negative margins cannot be attributed to label granularity alone. Moreover, PANDA contributed to `MOOZY`'s pretraining corpus (Table 2): its relative advantage should therefore be interpreted in light of potential pretraining–evaluation overlap.

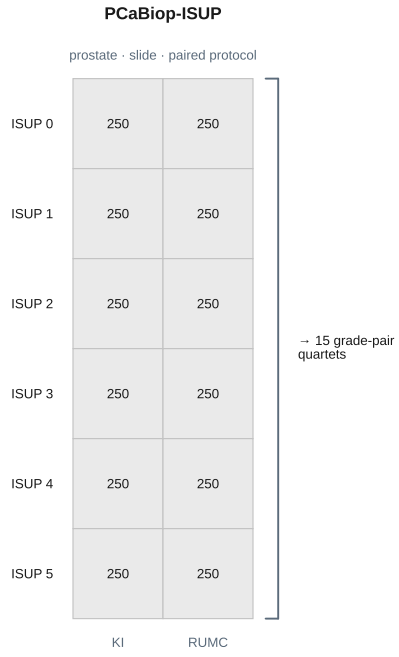


Figure 18: **PCaBiop-ISUP is balanced across ISUP grade and contributing centre.** Rows denote ISUP grade groups (grade 0 is benign, grades 1–5 increasingly aggressive prostate cancer) and columns the acquisition centre. Cells show the evaluated biopsies per ISUP grade×centre combination: 250 slides in each of the 6×2 cells, for 3,000 slides in total.

(a) Cancer detection									
Model	$k^\star$	bio bacc	conf bacc	RI	MaRI	CRoMa	$F(0)$	LTM_{10}	support
PRISM	3	0.968	0.992	**0.281**	**0.195**	**0.26**	**0.288**	**-0.39**	10.1%
MOOZY	9	**0.971**	0.999	0.236	0.181	-0.02	0.535	-0.41	24.3%
TITAN	9	0.915	1.000	0.015	0.001	-0.30	0.895	-0.60	49.1%
Prov-GigaPath	3	0.716	0.995	0.014	0.003	-0.41	0.990	-0.59	**63.9%**

(b) ISUP grading									
Model	$k^\star$	bio bacc	conf bacc	RI	MaRI	CRoMa	$F(0)$	LTM_{10}	support
MOOZY	11	**0.934**	0.999	**0.077**	**0.053**	**-0.09**	0.675	**-0.45**	**45.3%**
PRISM	3	0.858	0.990	0.063	0.039	-0.14	**0.651**	-0.51	33.4%
TITAN	1	0.804	1.000	0.001	0.000	-0.39	0.960	-0.61	19.6%
Prov-GigaPath	1	0.656	0.998	0.004	0.000	-0.47	0.993	-0.62	34.5%

Table 14: **Robustness of slide-level representations on PCaBiop and PCaBiop-ISUP.** The four slide-level encoders are ordered within each panel by CRoMa (m=5). Biological balanced accuracy (bio bacc) is reported at the model-specific, biologically selected $k^\star$; confounder balanced accuracy (conf bacc) is the maximum over the evaluated k grid. RI and MaRI are pooled at $k^\star$; support is the fraction of samples contributing to these fixed-k scores. $F(0)$ is the fraction with CRoMa < 0, and LTM_{10} is the mean of the lowest decile. Bold denotes the most favourable value in each score. **a,** Binary cancer detection on 1,000 slides. **b,** Six-class ISUP grading over 3,000 slides.