A 6M-Parameter Vision Transformer Trained from Scratch Is Competitive with 3–14× Larger Models for Mitotic-Figure Detection Across Tumour Types

Diego Garcia Rebate
VairoX Labs

© 2026 Diego Garcia Rebate. This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

Abstract

Vision transformers are usually reported to trail convolutional networks on mitotic-figure detection unless they are large and pretrained on external data. We ask whether a 6.05M-parameter ViT trained from scratch on 278 regions of MIDOG++ can compete on tumour types it has never seen. Karyo-T, a ViT-Tiny with a convolutional stem, 2D rotary positions, registers and a multi-level neck, is developed under a leave-one-domain-out protocol in which the held-out tumour never influences training, checkpoint or threshold, and is compared on one laptop against a same-size CNN, YOLO11m (20M) and a frozen 86M pathology foundation model (hibou-b). On three held-out tumour types it reaches micro F1 0.675 ± 0.013, 0.619 ± 0.008 and 0.809 ± 0.004 (3 seeds). After Holm correction over 92 comparisons it is ahead of the same-size CNN on canine lymphoma and shows no significant difference from YOLO11m or hibou-b on any fold, although the intervals are too wide to show equivalence within ± 0.03 F1. Two external tests of unseen tumour types give the same picture: random regions of canine mammary carcinoma slides (F1 0.722 for Karyo-T, 0.705 for YOLO11m, 0.708 for hibou-b) and human lung adenocarcinoma (recall 0.501, 0.433, 0.516). For every model the validation threshold transfers badly to the most shifted tumour, losing 0.04–0.21 F1 on lymphoma; one labelled region of the new tumour repairs most of it for Karyo-T.

Keywords mitotic-figure detection · domain generalisation · vision transformers · computational pathology · leave-one-domain-out evaluation

1 Introduction

Counting mitotic figures is part of tumour grading, and automating it is a long-standing benchmark of computational pathology [4, 6, 7]. Domain shift makes this hard, because a detector trained on some tumour types, scanners and laboratories must work on others. The MIDOG challenges and the MIDOG++ dataset [5] made this measurable. In the 2025 edition, YOLO-family and other convolutional detectors took most of the top places, ensembles helped, and model size did not correlate with rank [7]. Where transformers appear, they come with large pretrained backbones (DINOv2 in RF-DETR, pathology foundation models used frozen [9]); studies that compare plain ViTs with CNNs on mitoses find the ViT behind and attribute the gap to the CNN's inductive bias at small data [2].

We test the opposite hypothesis: that a small ViT, trained from scratch on MIDOG++ alone, can be competitive under domain shift if it is given the locality the task needs. Every design choice was taken on validation regions of the training tumours; the held-out tumour was used only for the final measurement; every reported number is a mean over three seeds with a paired bootstrap interval, corrected for the number of comparisons; and all compared models were trained on the same machine with the same scorer. Hypotheses were registered before their experiments, and failed ones are reported with the rest. Our claim is deliberately narrow: we find no significant difference between Karyo-T and models 3 and 14 times its size, and no evidence that it is better; our intervals are too wide to show that it is equivalent.

Contributions.

- Karyo-T, a 6.05M-parameter ViT point detector trained from scratch, and the controlled path that produced it: each locality prior and each training change screened on training-domain validation with a gate written in advance (Section 4).
- A leave-one-domain-out comparison on three tumour types against a same-size CNN, YOLO11m, and a frozen 86M pathology foundation model with the same neck and heads, all under one scorer, on one platform, with every training budget disclosed (Sections 5 and 6.1).
- Two external tests on tumour types absent from MIDOG++: MITOS_WSI_CMC [3], cut and registered before any model was scored and led by its random regions, and LUNG-MITO [22], a human tumour from another laboratory, registered before scoring and read for recall, with look-alike firing as a partial check on false alarms (Sections 6.3 and 6.4).
- Analyses of where the errors come from (a threshold that does not transfer, for every model), how sensitive the comparisons are to single regions and to test-time augmentation, how much a few labelled regions of the new tumour recover, and what deployment costs (Sections 6.5–6.10), plus two negative results on in-domain self-supervised pretraining (Section 6.11).

2 Related work

Mitosis detection under domain shift. MIDOG 2021 and 2022 winners predicted dense maps (segmentation or heatmaps) [4, 6]. In MIDOG 2025 the top single models were YOLO variants of about 20M parameters, the winner a 381M ensemble at F1 0.740, and real-time single-stage detectors such as RTMDet-S were competitive [7, 29]. MIDOG++ [5] provides leave-one-tumour-out reference numbers for a RetinaNet [27]. Published leave-one-domain-out results are hard to compare: several select checkpoints on the held-out domain [32], use a fixed threshold [11], report one run, or test on a different set; a recent systematic review concludes that cross-study comparability is limited by inconsistent evaluation and scarce cross-domain testing [1]. A leave-one-domain-out audit also showed that F1, calibration and count error disagree under shift, and that lymphoma is undercounted [11]; our models undercount lymphoma too.

Foundation models and transformers for mitoses. Frozen pathology foundation models used as detection encoders come close to an end-to-end ResNet-50 in-domain on MIDOG++ and do better out of domain on TUPAC16 [9]. For classification, LoRA [21] adaptation of foundation models nearly closes the out-of-domain gap [2]. Hibou-b [30], our foundation-model baseline, had no published mitosis detection result we could find. CellViT [20] pairs a large pretrained ViT with UNETR-style skips [14] for nuclei; we borrow the skip idea at 6M parameters from scratch.

Small ViTs. Convolutional stems [35], rotary position embeddings for images [17], register tokens [10] and simple feature pyramids for plain ViTs [26] are all published on natural images. We test them one at a time on a small-data detection task under domain shift, against the detectors and foundation models used for mitoses today.

3 Data and protocol

MIDOG++. 503 regions of 2 mm^2 from seven tumour types (human breast carcinoma, neuroendocrine tumour and melanoma; canine lung carcinoma, lymphoma, cutaneous mast cell tumour and soft-tissue sarcoma), five scanners, two species, with mitotic figures annotated as points and a non-exhaustive set of look-alikes ("imposters") [5]. We follow the authors' split file. For a held-out tumour, the model trains on the `train` cases of the other six (278–282 regions), validates on 20% of them chosen by case (71–72 regions), and is tested once on the held-out tumour's `test` cases (12, 12 and 11 regions; 840, 115 and 220 figures for lymphoma, neuroendocrine and melanoma).

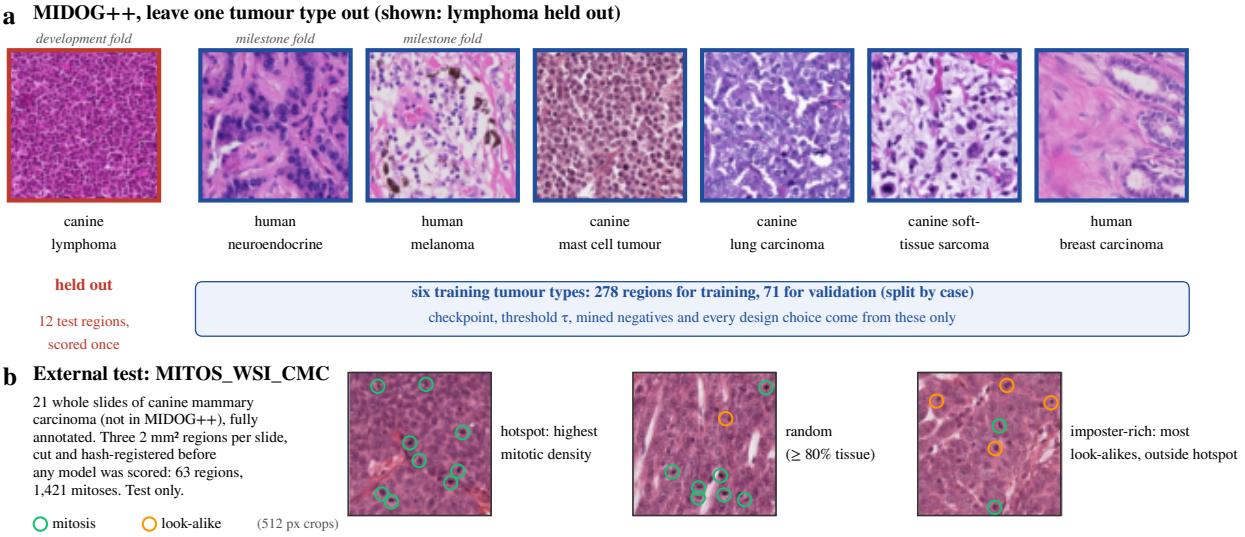

Figure 1. Evaluation protocol. (a) Leave one tumour type out on MIDOG++, shown for the development fold: the model trains and validates on regions of six tumour types and is scored once on the seventh. Every choice, including the detection threshold, is made on training-domain validation regions. (b) The external whole-slide test: three regions per slide of a tumour type absent from MIDOG++, chosen by fixed rules and registered before scoring. Real 512 px crops; circles mark annotated mitoses (green) and look-alikes (orange).

Folds. Canine lymphoma is the development fold (the largest shift); every design decision was taken on its training-domain validation. Neuroendocrine tumour (hardest) and melanoma (easiest) are milestone folds, run only with configurations already frozen and never used to decide. We did not run the other four tumour types (Section 8).

What the held-out tumour may touch. Only the final measurement: no pretraining, normalisation statistics, early stopping, checkpoint choice, threshold choice or hard-negative mining uses it. The code asserts this by case identifier.

Metrics. The primary metric is micro-averaged F1: detections are matched to annotated figures one-to-one within 7.5 μm (Hungarian matching, as in the organisers' code), TP, FP and FN are summed over all regions of the domain, then F1 is computed. The detection threshold τ is the one that maximises pooled F1 on training-domain validation and is then frozen. We also report average precision (AP), precision, recall and per-region count error. F1 at a threshold chosen on the test set is never reported as a result; we use it, labelled, as a diagnostic of threshold transfer (Section 6.6).

Statistics. Every configuration is trained with seeds 0, 1 and 2; we report mean ± sample standard deviation. Comparisons use a paired bootstrap (10,000 draws, seeds averaged within each arm before resampling) [13]: over test regions on MIDOG++, over slides on CMC. For each difference we report the 95% interval and a two-sided $p = 2\min(P(\Delta \leq 0), P(\Delta > 0))$, floored at $2/10{,}000$. All 92 comparisons the paper makes (Karyo-T against every baseline per fold, held out and on CMC by region type, the ensemble, and the registered tests H2 and H6) form one family, and we apply Holm's correction to it [19]; Appendix Table 9 lists every one. We call a model *ahead* only when the Holm-adjusted p is below 0.05; otherwise we say the difference is not significant, which is not the same as equivalent. With 92 comparisons the smallest attainable adjusted p is 0.018. To test equivalence we use two one-sided tests on the same bootstrap draws, with a margin of ±0.03 F1 that we set after the results were known, corrected as their own family of 39 (Appendix Table 12). The test-time augmentation comparisons were added later and are corrected as their own family (Table 10). Analyses added after the main results were known are marked *post hoc* in their section titles and captions. Screening runs during development use one seed and are reported only as development decisions.

External whole-slide test. MITOS_WSI_CMC [3] has 21 completely annotated whole slides of canine mammary carcinoma, a tumour type absent from MIDOG++. Before any model saw them, we cut three 2 mm^2 regions per slide inside the tumour annotation, with at least 80% tissue: a random region, the region of highest mitotic density (hotspot), and the region of highest look-alike density outside the hotspot (imposter-rich). The 63 regions contain 1,421 mitoses (198 random, 833 hotspot, 390 imposter-rich) and were registered by hash before scoring. Random regions follow no selection rule of ours beyond tissue and tumour, so we lead with them; hotspot and imposter-rich regions are reported after. Each model is applied at its own fold's validation threshold. Ten of the 14 MIDOG 2025 teams trained on this dataset [7]; we use it only for testing.

Second external test: LUNG-MITO. LUNG-MITO [22] contains 1024 px tiles of human lung adenocarcinoma from the University of Szeged, scanned on a 3DHISTECH scanner and annotated by mitotic phase, with look-alikes as a separate class. We inspected its metadata, never its images, after the MIDOG++ and CMC results were known. Three facts fixed how we use it. The tiles were selected around mitoses (91% contain one), so precision and F1 would be inflated; there are no patient or slide identifiers, so no clustered interval is possible; and 1,456 tiles have no stated size or resolution. Before scoring any model, we froze the 2,070 tiles at 0.24 μm/px (2,385 mitoses: prophase to telophase and atypical figures; look-alikes excluded) and registered the manifest by hash. We report only recall at each model's own validation threshold, the mean over seeds, without an interval.

Hardware. Every model in this paper, including the baselines and the foundation-model reference, was trained on one MacBook Pro (Apple M5, 24 GB unified memory) with PyTorch on Metal; every run records its platform.

4 Model and its development

4.1 Karyo-T

Karyo-T is a ViT-Tiny [12] (width 192, 12 blocks, 3 heads, MLP ratio 4) turned into a CenterNet-style point detector [36] (Figure 2):

- **Convolutional stem** [35]: four 3×3 stride-2 convolutions (widths 24–192), each followed by LayerNorm over channels and GELU, then a 1×1 projection; tokens at stride 16. We use LayerNorm so that the model stores no statistics of the training domains, as BatchNorm would.
- **2D axial rotary position embedding** [17] in every attention layer, half of each head's dimensions rotated by column and half by row; no absolute position table.
- **Four register tokens** [10] and no class token.
- **Multi-level neck**: patch tokens after blocks 4, 8 and 12 are projected and summed at stride 16, upsampled to stride 8 by a transposed convolution, and fused with the stem's stride-8 map (a UNETR-style skip [14]).
- **Heads** at stride 8: a heatmap trained with the CenterNet focal loss [25, 36] on Gaussian targets (σ = 16 px) and a sub-cell offset trained with L1. Peaks are 3×3 local maxima above τ.

The detector has 6,045,187 parameters. Whole regions are processed in 512×512 windows with 64 px overlap; a detection is kept only if it lies in the window's inner part, and detections closer than 15 px are merged.

4.2 Training

Windows of 512 px are drawn on the fly, centred near a random mitosis, a random look-alike or a random tissue point, a third each. From epoch 20, a quarter of the windows are instead centred on the model's own false positives on the training cases (score above 0.3), re-mined every 10 epochs. Augmentation, drawn per window: one of the 8 flips and 90° rotations of the square (the labelled points move with it); HED stain jitter [34], each stain channel scaled by

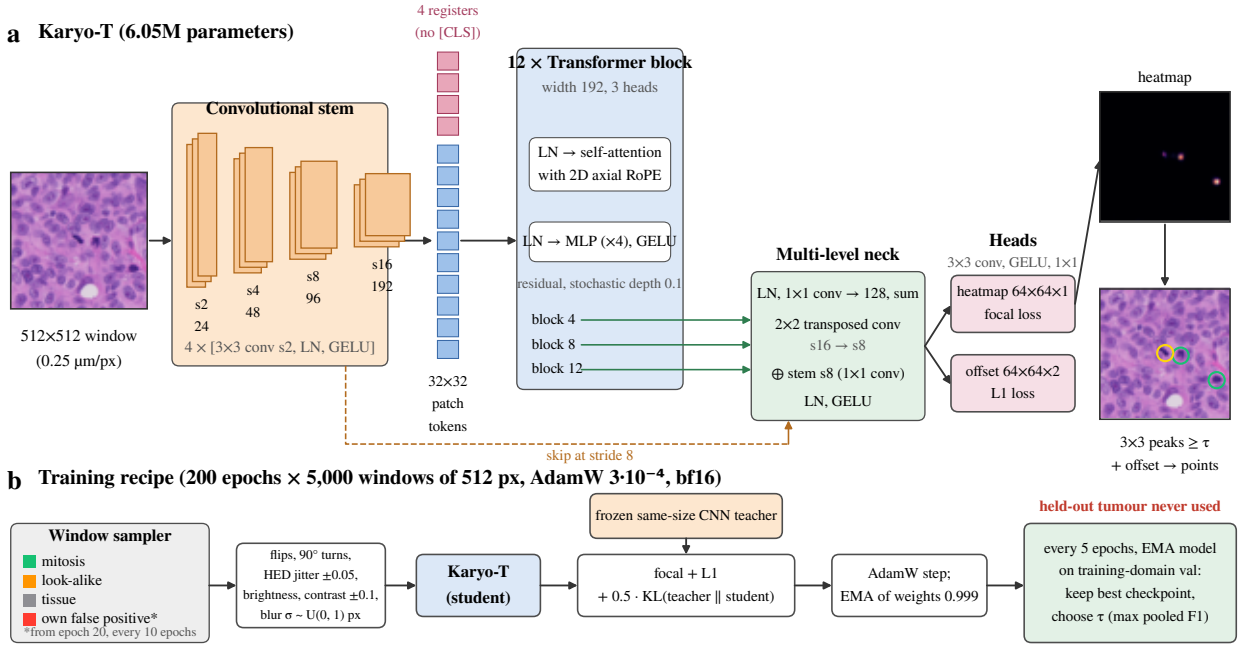

Figure 2. (a) Karyo-T. A convolutional stem turns a 512 px window into 32×32 patch tokens; four register tokens replace the class token; twelve transformer blocks use 2D rotary positions; a neck fuses blocks 4, 8 and 12 with the stem's stride-8 map; two heads predict a heatmap of mitosis centres and a sub-cell offset at stride 8. Shown on a real held-out melanoma window: the heatmap and the detections at the validation threshold (green: detected figure; yellow: missed). (b) The training recipe, which never uses the held-out tumour.

$U(0.95, 1.05)$ and shifted by $U(-0.05, 0.05)$; brightness and contrast jitter of ± 0.1 (a shift of $U(-0.1, 0.1)$ on intensities in $[0, 1]$ and a scaling by $U(0.9, 1.1)$ about the window mean); and a Gaussian blur with $\sigma \sim U(0, 1)$ px; the result is clipped to $[0, 1]$. Every model we trained with our own code (both CNNs, the standard ViT, the architecture-only Karyo-T and hibou-b) uses this augmentation; YOLO11m uses its own. We train for 200 epochs of 5,000 windows (batch 32) with AdamW [28] (lr $3 \cdot 10^{-4}$, weight decay 0.05, 3 warm-up epochs, cosine decay, gradient clipping 1.0, stochastic depth 0.1), bf16 autocast and `torch.compile`, and keep an exponential moving average of the weights (decay 0.999) [33]. The student also matches the heatmap of a same-size CNN detector trained on the same fold (output distillation [18], a Bernoulli KL on heatmap probabilities with weight 0.5); that CNN is the "same-size CNN" baseline below, so Section 6.2 separates what it contributes. Every 5 epochs the EMA model is run on all validation regions; the checkpoint with the best validation score ($0.6\,\mathrm{F1} + 0.3\,\mathrm{FROC} + 0.1\,\mathrm{AP}$, each averaged over the six training tumours) is kept, and τ is chosen on the same validation set.

4.3 How each piece was chosen

Figure 3 shows every change we screened. Each was tried on top of the last kept change, with one seed on the development fold, against a gate on training-domain validation F1 written before the run. Changes meant to help on unseen domains (stain and scale augmentation), which training-domain validation cannot measure, were screened with a second tumour (canine lung carcinoma) also left out of training and validation and scored as an inner held-out domain. Held-out lymphoma numbers were logged but never used to decide.

The architecture stage carried most of the improvement; the convolutional stem alone moved validation F1 by +0.10. Every stronger stain augmentation (HED jitter of ± 0.15, RandStainNA) and scale augmentation we tried lowered the score on the inner held-out tumour, so the final model keeps the augmentation it started with: mild HED jitter, brightness, contrast and blur.



Figure 3. Development path on the lymphoma fold. Each screen is one seed (a decision, not a result) and was tried on top of the last kept change against a gate on training-domain validation F1 written before the run; the starting reference is the mean of 3 seeds. In stage 2, canine lung carcinoma was also left out and scored as an inner held-out domain (triangles), because stain and scale augmentation are meant to help on unseen domains, which training-domain validation cannot show. The kept configuration reached validation F1 0.779 ± 0.006 over 3 seeds.

5 Baselines and training budgets

All baselines are trained on the same folds, validated and thresholded the same way, and scored by the same code. Table 1 gives each one's training budget.

- **ViT, standard recipe**: ViT-Tiny/16 with a class token and learned absolute positions, the same neck type and heads (6.19M); the starting point of Figure 3.
- **Same-size CNN**: a ResNet [15] with Karyo-T's heads at stride 8 (6.06M), 384 px windows, 60 epochs. It is also Karyo-T's distillation teacher. On lymphoma we also trained it at Karyo-T's full budget (512 px, 200 epochs, mining, EMA, no teacher; "same budget").
- **YOLO11m** [24] (20.05M), 20 epochs of 15,000 windows drawn like Karyo-T's, a 64 px box around each point, Ultralytics defaults plus vertical flips; box centres are scored as points with the same tiling, threshold rule and scorer.
- **Frozen hibou-b** [30], a DINOv2 [31] ViT-B/14 with registers pretrained on external pathology slides (86.4M, 0.66M trained), with Karyo-T's multi-level neck, heads and full training recipe, including distillation from the same CNN; only the backbone differs. Its backbone was pretrained on data we do not control, so we report it as a reference on a separate external-weights track.
- **RetinaNet** numbers from the MIDOG++ paper [5] (copied; one run each, trained under its own protocol).

YOLO11m was trained without COCO weights, by choice. Ultralytics starts YOLO11m from COCO-pretrained weights by default. Our main track allows only weights learned from MIDOG++ training-domain pixels, and every model on it, Karyo-T included, starts from random initialisation; external pretraining is confined to the hibou-b reference. We therefore trained YOLO11m from scratch (`yolo11m.yaml`). A default user would start from COCO, and we did not measure how much that would add, so our YOLO11m may be weaker than the one most readers have in mind. It also saw 0.30M training windows against Karyo-T's 1.00M (Table 1). We did not rerun it at a larger budget; only the CNN got Karyo-T's budget, on lymphoma.

Table 1. Training budget of every model, per run, on the Apple M5. Windows seen = epochs × windows per epoch (YOLO11m: a fixed set of 15,000 windows, 20 passes). Hours: wall-clock training time, mean over runs; "+ 5.9" is the CNN teacher, trained once per fold and seed and reused. Pretrained: weights not learned from MIDOG++ training domains (hibou-b: backbone, frozen; 0.66M parameters trained).

Model	Params (M)	Pretrained	Teacher	Window (px)	Epochs	Windows seen	Hours	Runs
ViT, standard recipe	6.19	none	–	384	60	0.30M	2.0	9
Same-size CNN	6.06	none	–	384	60	0.30M	5.9	9
CNN, same budget	6.06	none	–	512	200	1.00M	37.6	3
YOLO11m	20.05	none (no COCO)	–	512	20	0.30M	8.9	9
Frozen hibou-b	86.4 (0.66)	external, frozen	CNN	512	200	1.00M	24.8 + 5.9	9
Karyo-T, architecture only	6.05	none	–	384	120	0.60M	5.1	9
Karyo-T	6.05	none	CNN	512	200	1.00M	24.4 + 5.9	9

Table 2. Held-out micro F1 (mean ± sd over 3 seeds) and AP on MIDOG++, each model at its own training-domain validation threshold. Parameters in millions. Bold: best in each column; underlined: second best. Differences between Karyo-T, YOLO11m and hibou-b are not significant after Holm correction (Table 3). †One ensemble of the three Karyo-T seeds (heatmaps averaged before peak finding), so one run per fold. ‡Copied from [5].

		Lymphoma		Neuroendocrine		Melanoma		Mean
Model	Params	F1	AP	F1	AP	F1	AP	F1
RetinaNet‡		0.57		0.59		0.79		0.650
ViT, standard recipe	6.2	0.484 ± 0.034	0.615	0.372 ± 0.025	0.419	0.584 ± 0.020	0.567	0.480
Same-size CNN	6.1	0.547 ± 0.030	0.816	0.564 ± 0.035	0.639	0.774 ± 0.023	0.840	0.628
same budget	6.1	0.625 ± 0.015	**0.842**					
YOLO11m	20.1	0.652 ± 0.018	0.764	0.598 ± 0.007	0.655	0.788 ± 0.022	0.850	0.679
Frozen hibou-b	86.4	**0.699** ± 0.031	0.837	**0.641** ± 0.007	**0.706**	0.813 ± 0.004	0.863	**0.718**
Karyo-T, architecture only	6.05	0.657 ± 0.002	0.751	0.561 ± 0.029	0.617	0.743 ± 0.017	0.800	0.654
Karyo-T	6.05	0.675 ± 0.013	0.816	0.619 ± 0.008	0.675	0.809 ± 0.004	0.873	0.701
Karyo-T×3†	18.1	0.659	0.828	0.621	0.684	**0.825**	**0.879**	0.702

6 Results

6.1 Held-out tumour types

The locality priors close the gap between ViT and CNN. Table 2, Table 3 and Figure 4a,c give the held-out results. With a standard recipe, the 6M ViT is far behind the same-size CNN on every fold (0.480 vs 0.628 mean F1), as the literature predicts. The architecture changes alone (Figure 3, stage 1) lift it to 0.654. Our registered test of this (H2) asked whether the new architecture closes at least half of the gap between the standard ViT and the CNN; the statistic $D = F_1(\text{Karyo-T}) - \frac{1}{2}[F_1(\text{ViT}) + F_1(\text{CNN})]$ was +0.141 [+0.098, +0.191] on lymphoma (Holm p 0.018, supported), +0.064 [+0.031, +0.102] on melanoma (Holm p 0.076) and +0.093 [−0.047, +0.146] on neuroendocrine (not established: one region holds 55% of that fold's figures). After correction, H2 is supported on lymphoma only.

Against the baselines. With the training changes, Karyo-T is ahead of the same-size CNN on lymphoma (+0.128, Holm p 0.018) and stays ahead when the CNN gets Karyo-T's full training budget (+0.050, Holm p 0.018). On neuroendocrine and melanoma the difference from the CNN is not significant (+0.055 and +0.034, intervals touching or containing zero). The budget-matched CNN ranks detections better on lymphoma (AP 0.842 vs 0.816), so Karyo-T's F1 lead there is at the operating point, not in ranking (Section 6.6). The difference from YOLO11m is not significant on any fold: Karyo-T is ahead by about 0.02 F1 at each point estimate, and every Holm-adjusted p is 1.00. Nor is the difference from frozen hibou-b; hibou-b is ahead at each point estimate (0.005–0.024; on lymphoma the interval is [−0.050, +0.000]), again with Holm-adjusted p of 1.00. Neither difference is shown to be small. No comparison with YOLO11m or hibou-b is equivalent within ±0.03 F1 (post hoc, Table 12): held out, their 90% intervals reach as far as +0.096 and −0.058, and even before correction only four CMC comparisons over all regions fall inside the margin.

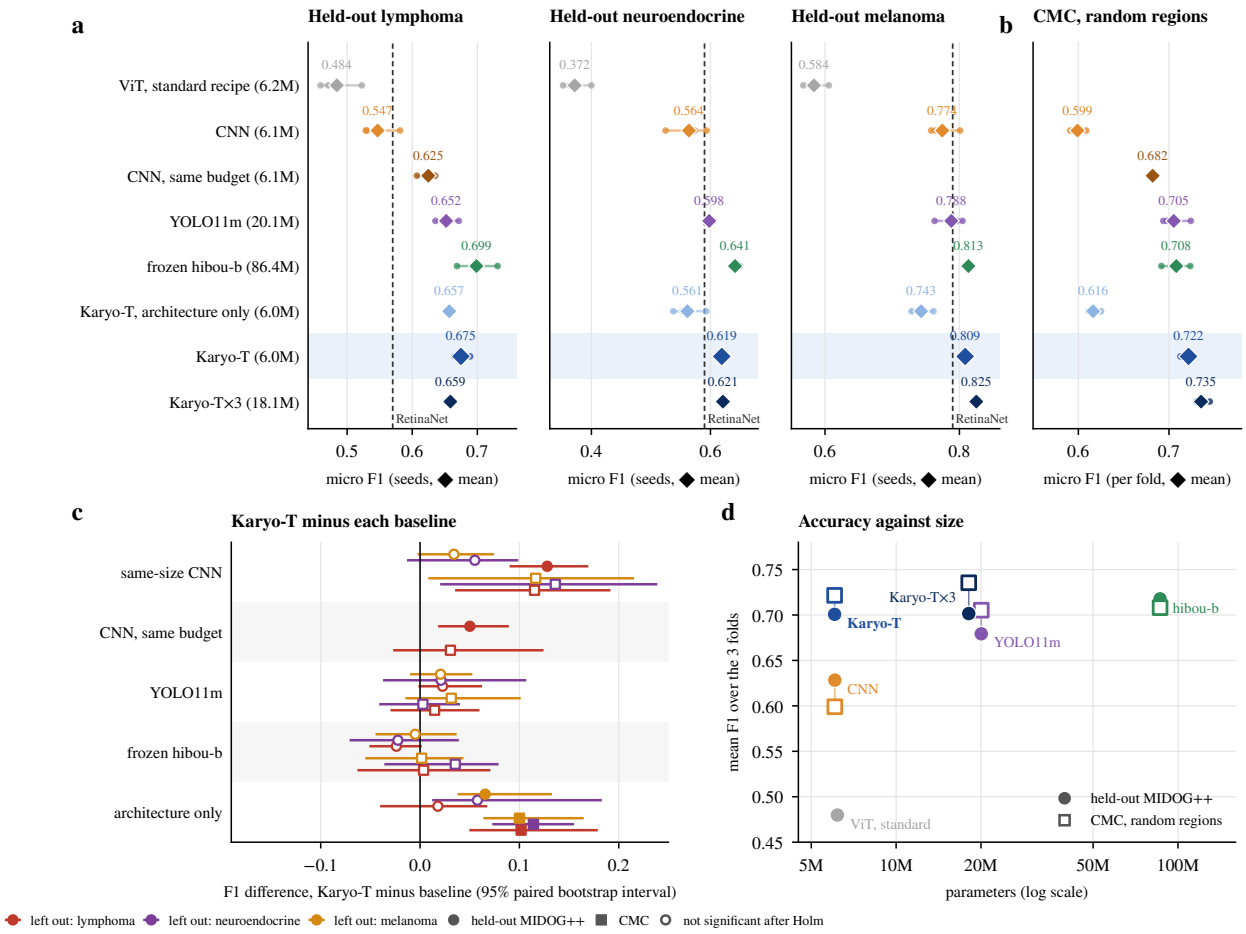

Figure 4. Main results. (a) Held-out micro F1 per left-out tumour type: small dots are seeds, diamonds their mean; the dashed line is the published RetinaNet [5]. (b) The same models on the random regions of the MITOS_WSI_CMC whole-slide test; small dots are the means of the models trained without each of the three tumour types. (c) Paired bootstrap of the F1 difference between Karyo-T and each baseline, per fold, on held-out MIDOG++ (circles) and on CMC random regions (squares); filled markers stay significant after Holm correction over all 92 comparisons. (d) Mean F1 over the three folds against parameter count, held out (circles) and on CMC random regions (squares).

Karyo-T is above the published RetinaNet on all three folds, although that comparison is not paired and the protocols differ.

Ensembles. Averaging three seeds helps on melanoma (+0.017, Holm p 0.018) and lowers F1 on lymphoma (−0.016, Holm p 0.14), where it ranks better (AP 0.828 vs 0.816) but its averaged scores push the validation threshold too high.

6.2 How much of the gain over the CNN is the CNN teaching Karyo-T (post hoc)

Karyo-T is distilled from the same-size CNN it is compared with, so part of its lead over that CNN could be the teacher's knowledge passed on. Our runs separate the architecture from the training stage, not distillation from the other training changes.

The architecture alone, with no teacher ("architecture only": 384 px windows, 120 epochs), is ahead of the CNN only on held-out lymphoma (+0.110 [+0.034, +0.189]). Its difference from the CNN is not significant on neuroendocrine (−0.003 [−0.173, +0.055]), on melanoma, where it is behind at the point estimate (−0.031 [−0.102, +0.008]), or on CMC on any fold (+0.007 to +0.024, all intervals containing zero). Everything else in Karyo-T's lead over the CNN

Table 3. Karyo-T minus each baseline: difference in micro F1, 95% paired bootstrap interval and Holm-adjusted p (family of 92; the smallest attainable value is 0.018). Regions resampled on MIDOG++, slides on CMC. Bold: Holm-adjusted $p < 0.05$. Hotspot, imposter-rich and all-region CMC comparisons are in Appendix Table 9.

Karyo-T minus	Lymphoma	Neuroendocrine	Melanoma
Held-out MIDOG++			
same-size CNN	**+0.128 [+0.091, +0.168] 0.018**	+0.055 [−0.012, +0.098] 1.00	+0.034 [−0.001, +0.073] 1.00
same-size CNN, same budget	**+0.050 [+0.019, +0.088] 0.018**		
YOLO11m	+0.023 [−0.000, +0.061] 1.00	+0.021 [−0.036, +0.106] 1.00	+0.021 [−0.009, +0.051] 1.00
frozen hibou-b	−0.024 [−0.050, +0.000] 1.00	−0.022 [−0.070, +0.038] 1.00	−0.005 [−0.044, +0.036] 1.00
Karyo-T, architecture only	+0.018 [−0.039, +0.066] 1.00	+0.058 [+0.013, +0.182] 0.86	**+0.066 [+0.039, +0.132] 0.018**
CMC, random regions			
same-size CNN	+0.115 [+0.037, +0.190] 0.51	+0.136 [+0.022, +0.238] 0.89	+0.116 [+0.010, +0.214] 1.00
same-size CNN, same budget	+0.031 [−0.026, +0.123] 1.00		
YOLO11m	+0.015 [−0.028, +0.059] 1.00	+0.003 [−0.040, +0.039] 1.00	+0.031 [−0.013, +0.100] 1.00
frozen hibou-b	+0.004 [−0.062, +0.070] 1.00	+0.035 [−0.035, +0.078] 1.00	+0.002 [−0.054, +0.042] 1.00
Karyo-T, architecture only	**+0.102 [+0.051, +0.178] 0.018**	**+0.114 [+0.074, +0.154] 0.018**	**+0.100 [+0.065, +0.163] 0.018**

Table 4. MITOS_WSI_CMC (21 slides, 63 regions, test only), mean over the three left-out folds and 3 seeds. Random regions first: they follow no selection rule of ours. Then F1 by left-out fold on all regions, hotspot and imposter-rich F1, imposter-rich precision and AP (all regions). Bold: best in each column; underlined: second best. No difference between Karyo-T, YOLO11m and hibou-b on random regions is significant after Holm correction. *Lymphoma fold only. †One ensemble per fold.

	Random	F1, all regions				F1 by region type		Imposter-rich	
Model	F1	lymph.	neuro.	melan.	mean	hotspot	imposter	precision	AP
Same-size CNN	0.599	0.634	0.664	0.670	0.656	0.676	0.648	0.684	0.721
same budget*	0.682	0.701				0.714	0.686	0.719	0.768
YOLO11m	0.705	0.722	0.744	0.738	0.735	<u>0.760</u>	0.703	0.672	0.796
Frozen hibou-b	0.708	0.723	0.737	0.743	0.734	0.753	0.711	0.700	0.796
Karyo-T, architecture only	0.616	0.658	0.671	0.680	0.670	0.689	0.660	0.668	0.726
Karyo-T	<u>0.722</u>	<u>0.728</u>	<u>0.754</u>	<u>0.755</u>	<u>0.745</u>	0.759	<u>0.731</u>	<u>0.743</u>	<u>0.809</u>
Karyo-T×3†	**0.735**	**0.733**	**0.765**	**0.763**	**0.754**	**0.763**	**0.746**	**0.775**	**0.823**

appears in the training stage, which adds five changes at once: distillation from that CNN, mined hard negatives, weight averaging, 512 px windows and 200 epochs. The stage adds +0.018, +0.058 and +0.066 held out and +0.070 to +0.083 on CMC (Table 9). We did not train a 3-seed Karyo-T without the teacher, so its share of that stage cannot be stated with an interval. The only isolated measurement is a one-seed screen, in which distillation added +0.012 validation F1 of the stage's +0.035 at screening budget (Figure 3).

Read plainly: on neuroendocrine, on melanoma and on CMC, Karyo-T's entire lead over the CNN comes from a training stage that includes the CNN as teacher, and up to all of it could be owed to distillation. Two results bound this. The teacher itself scores 0.547 on held-out lymphoma, well below its student's 0.675, so the student is not copying a better model. And a CNN trained at Karyo-T's full budget without a teacher reaches 0.625, still behind Karyo-T (+0.050). Frozen hibou-b is distilled from the same CNN, so the comparisons with hibou-b are like for like in this respect; YOLO11m has no teacher.

6.3 Whole-slide test on an unseen tumour type

Random regions. On the 21 random regions, the least arguable part of the test, Karyo-T scores 0.722 F1, YOLO11m 0.705 and frozen hibou-b 0.708 (Table 4, Figure 4b). Slide-bootstrap differences, Karyo-T minus YOLO11m, are +0.015, +0.003 and +0.031 by left-out fold, and minus hibou-b +0.004, +0.035 and +0.002; every interval contains zero, so neither difference is significant. Against the same-size CNN the differences are large (+0.115, +0.136, +0.116) and each uncorrected interval excludes zero, but with only 21 regions none survives Holm correction (adjusted p 0.51–1.00). The training stage is clearly ahead of the architecture alone here (+0.100 to +0.114, Holm p 0.018 on every fold).

Table 5. LUNG-MITO (2,070 tiles, 2,385 mitoses, human lung adenocarcinoma, another laboratory; test only). Recall at each model's own validation threshold: mean over the three left-out folds and 3 seeds, and the range of the fold means (registered before scoring). Tiles were selected around mitoses and carry no patient identifiers, so precision, F1 and intervals are not reported. Post hoc, added after scoring as a partial check on false alarms: the share of the 515 annotated look-alikes (none within reach of a mitosis) that draw a detection, detections per tile on the 293 tiles without an annotated mitosis, and recall at a matched level of each (look-alike firing 0.1; 0.5 detections per mitosis-free tile), which removes the threshold.

	Recall at val τ		False alarms at val τ		Recall at matched false alarms	
Model	mean	fold range	look-alikes fired	per mitosis-free tile	look-alikes 0.1	0.5 per tile
Same-size CNN	0.399	0.353–0.443	0.015	0.144	0.745	0.594
same budget	0.389		0.012	0.100	0.740	0.669
YOLO11m	0.433	0.405–0.479	0.024	0.186	0.694	0.630
Frozen hibou-b	0.516	0.501–0.529	0.029	0.096	0.745	0.730
Karyo-T, architecture only	0.454	0.445–0.462	0.034	0.154	0.638	0.611
Karyo-T	0.501	0.482–0.516	0.017	0.146	0.732	0.667
Karyo-T×3	0.498	0.478–0.525	0.017	0.132	0.745	0.680

All regions, hotspots and look-alikes. Over all 63 regions the picture is the same. The differences from YOLO11m (+0.006, +0.010, +0.016) and from hibou-b (+0.004, +0.018, +0.012) are not significant after correction, and Karyo-T is ahead of the same-size CNN on every fold (+0.084 to +0.094, Holm p 0.018); the difference from the budget-matched CNN is not significant (+0.026, Holm p 1.00), and neither is any hotspot difference from the two larger models (−0.009 to +0.014). In imposter-rich regions Karyo-T has the highest precision among single models (0.743, against 0.672 for YOLO11m and 0.700 for hibou-b), and one comparison with a larger model survives correction: Karyo-T minus YOLO11m on the melanoma fold, +0.037 [+0.017, +0.055], Holm p 0.038. We do not read one fold of one region type as a general lead. Per-region mitotic counts from Karyo-T correlate with the truth at $r = 0.97$–0.98 (mean absolute error 5.0–6.7 figures per region), against $r = 0.81$–0.86 for the CNN.

6.4 Second external test: human lung adenocarcinoma

LUNG-MITO is a different tumour type, laboratory and scanner from anything in MIDOG++ or CMC. Table 5 gives recall at each model's validation threshold. The order matches the other two test sets: frozen hibou-b recovers 0.516 of the figures, Karyo-T 0.501, YOLO11m 0.433 and the same-size CNN 0.399. The gap between Karyo-T and hibou-b is smaller than the spread between the folds of either model.

False alarms (post hoc). Recall alone would favour a model that simply fires more, so after scoring we added two partial measures of false alarms that the set allows: how many of the annotated look-alikes draw a detection, and how many detections fall on the tiles without an annotated mitosis. Neither is a precision, since look-alikes are annotated selectively and we do not know whether the mitosis-free tiles were checked exhaustively. At its own threshold, Karyo-T fires on fewer look-alikes than YOLO11m (0.017 vs 0.024) and makes fewer detections per mitosis-free tile (0.146 vs 0.186), so its higher recall does not come from firing more. Matching the false alarms instead of using each model's threshold changes the picture. At the same look-alike firing rate (0.1), hibou-b, the same-size CNN and Karyo-T×3 recover 0.745, the budget-matched CNN 0.740, Karyo-T 0.732 and YOLO11m 0.694; at rates 0.2 and 0.3 the pattern is the same, with Karyo-T 0.01–0.02 below the CNNs, hibou-b and Karyo-T×3, and YOLO11m further behind. At 0.5 detections per mitosis-free tile, hibou-b leads (0.730), followed by Karyo-T×3 (0.680), the budget-matched CNN and Karyo-T (0.669 and 0.667), YOLO11m (0.630) and the same-size CNN (0.594); at 0.25 and 1.0 detections per tile hibou-b still leads, and Karyo-T stays within 0.015 of the budget-matched CNN and ahead of YOLO11m and the same-size CNN. Without intervals we read these as orderings, not differences: at a matched operating point Karyo-T is close to the CNNs, ahead of YOLO11m and behind hibou-b, so most of its recall lead over the CNNs at the validation threshold is again a matter of the threshold, as on lymphoma.

Recall is low for every model because the scores are shifted down on this tumour, as on lymphoma (Section 6.6): at a threshold of 0.1, Karyo-T would recover 0.81 and hibou-b 0.84, at a precision we cannot measure here. By phase

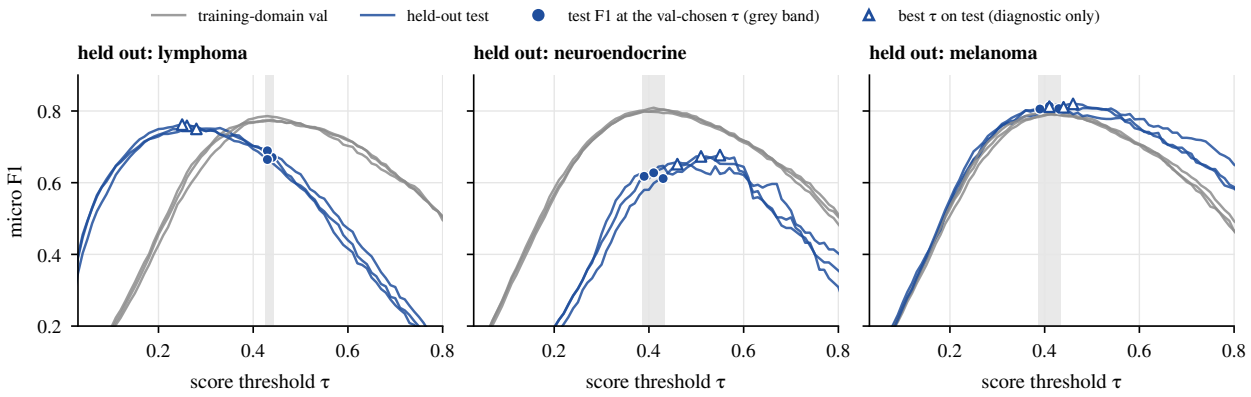

Figure 5. F1 against the score threshold for the three Karyo-T seeds, on training-domain validation (grey) and on the held-out tumour (blue). Dots: test F1 at the validation-chosen threshold, the reported result. Triangles: the best threshold on the test set, shown as a diagnostic and never reported as a result. On lymphoma the held-out scores are shifted down and the validation threshold falls on the declining side of the held-out curve; on melanoma the two curves coincide. Table 6 gives the same diagnostic for every model.

Table 6. Post hoc. Threshold transfer, for every model: held-out F1 at the validation-chosen threshold → F1 at the best threshold on the test set (a diagnostic, never a result), mean over seeds, and the gain in brackets. CMC: range of the gain over the three left-out folds.

Model	Lymphoma	Neuroendocrine	Melanoma	CMC
ViT, standard recipe	0.484 → 0.595 (+0.111)	0.372 → 0.457 (+0.085)	0.584 → 0.593 (+0.010)	
Same-size CNN	0.547 → 0.760 (+0.213)	0.564 → 0.644 (+0.080)	0.774 → 0.782 (+0.008)	+0.008 to +0.028
same budget	0.625 → 0.770 (+0.146)			+0.009
YOLO11m	0.652 → 0.733 (+0.080)	0.598 → 0.645 (+0.047)	0.788 → 0.793 (+0.005)	+0.002 to +0.010
Frozen hibou-b	0.699 → 0.779 (+0.080)	0.641 → 0.698 (+0.057)	0.813 → 0.818 (+0.005)	+0.003 to +0.008
Karyo-T, architecture only	0.657 → 0.697 (+0.040)	0.561 → 0.623 (+0.062)	0.743 → 0.757 (+0.014)	+0.004 to +0.013
Karyo-T	0.675 → 0.757 (+0.082)	0.619 → 0.667 (+0.048)	0.809 → 0.814 (+0.006)	+0.001 to +0.021
Karyo-T×3	0.659 → 0.766 (+0.108)	0.621 → 0.670 (+0.049)	0.825 → 0.825 (+0.000)	+0.000 to +0.025

(post hoc), every model finds prophase figures least often (Karyo-T 0.24 of 485), prometaphase to telophase more often (0.50 of 1,191), and atypical figures most often (0.68 of 709). Early prophase is often not counted as a mitotic figure in MIDOG++-style labelling, which may explain part of the first number.

6.5 Sensitivity to single regions (post hoc)

Each held-out fold has only 11–12 test regions, and on neuroendocrine one region (case 353) holds 55% of the figures. We re-scored the saved detections with each test region left out in turn. Karyo-T minus YOLO11m stays between +0.008 and +0.027 on lymphoma and between +0.013 and +0.033 on melanoma, and varies from −0.003 to +0.068 on neuroendocrine. Karyo-T minus hibou-b stays negative or near zero on every fold (−0.037 to +0.003), and Karyo-T minus the CNN stays positive (+0.024 to +0.142). No single region decides any comparison except Karyo-T against YOLO11m on neuroendocrine, whose sign depends on one region.

Without case 353, neuroendocrine F1 falls for every model, because that region carries most of the true positives: Karyo-T 0.402, hibou-b 0.406, the CNN 0.361 and YOLO11m 0.335. On the remaining 11 regions Karyo-T minus YOLO11m is +0.068 [−0.032, +0.119], minus the CNN +0.041 [−0.029, +0.116], and minus hibou-b −0.003 [−0.098, +0.050]; the conclusions of Table 3 do not change.

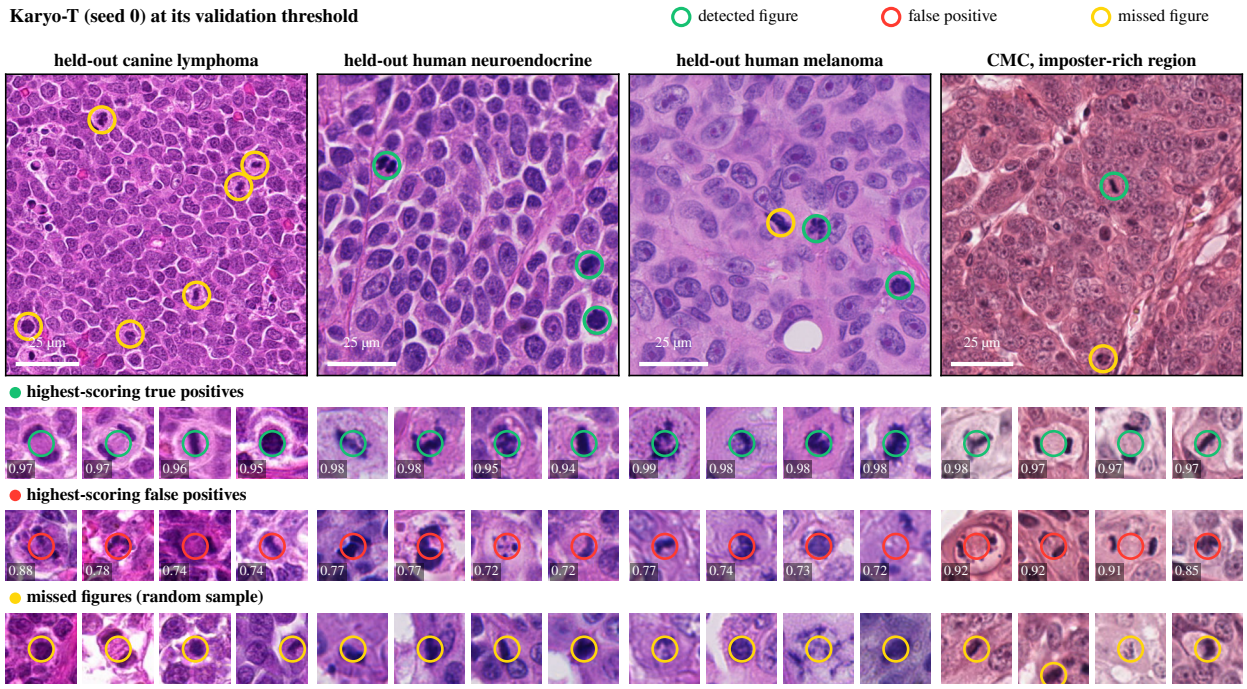

Figure 6. Karyo-T (seed 0) on real held-out regions, at its validation threshold. Top: for each held-out tumour, the 512 px window of its test regions with the most annotated figures, and the CMC imposter-rich region with the most look-alikes (lymphoma-fold model), selected by that rule. Below: over all of the domain's test regions, the four highest-scoring true and false positives (with scores) and four missed figures drawn at random. On lymphoma most figures in the window score just below the threshold (yellow), the under-detection of Figure 5. Several high-scoring false positives look mitotic but are labelled look-alikes or unlabelled.

6.6 Where the errors come from: thresholds that do not transfer

Every held-out F1 measures two things at once: how well the model separates mitoses from background on the new tumour, and whether its scores land where its threshold expects them. Figure 5 separates the two for Karyo-T, and Table 6 does it for every model. The problem is general. On lymphoma every model loses 0.04–0.21 F1 to the threshold, on neuroendocrine 0.05–0.09, on melanoma at most 0.014 and on CMC at most 0.028. On lymphoma all thresholds are too strict (best τ 0.10–0.36 against validation 0.37–0.54); on neuroendocrine all are too loose. On the training tumours the best threshold gains at most 0.017 F1.

The diagnostic also changes how the comparisons above should be read. At its own best threshold on lymphoma, the CNN reaches 0.760 and Karyo-T 0.757, so Karyo-T's lead over the CNN on lymphoma is entirely in how well its threshold transfers, not in how well it separates mitoses. The CNN loses the most to the threshold (0.213), Karyo-T 0.082, YOLO11m and hibou-b 0.080. Lymphoma is under-confident (detections scoring 0.2–0.3 are 34% true there, at most 15% on the other tumours), neuroendocrine over-confident, melanoma and CMC close to calibrated. The error counts agree: on lymphoma Karyo-T makes 58 false positives and 387 false negatives at its threshold, so it undercounts, as reported for RetinaNet [11]. On neuroendocrine one region holds 33 of 64 false positives. Figure 6 shows examples. Reading the 20 top-scoring false positives per domain by eye (the author, not a pathologist), between 5 and 12 look mitotic but are labelled look-alikes or are unlabelled; the others are mostly dense, pyknotic or fragmented nuclei, second detections on anaphase pairs, and (on melanoma) pigment-laden cells.

6.7 Test-time augmentation (post hoc)

We applied test-time augmentation to every compared model: each window is run in its 8 dihedral orientations (four rotations, with and without a flip), the detections are mapped back, and detections from different views within 15 px are fused into one, at the score-weighted mean position, with the sum of view scores divided by 8 as its score. Because

Table 7. Post hoc. Test-time augmentation: micro F1 without / with TTA (8 dihedral views, detections fused; threshold re-chosen with TTA on training-domain validation), mean over seeds. CMC: mean over the three left-out folds. Comparisons and Holm-adjusted p in Appendix Table 10.

	Held-out MIDOG++			CMC	
Model	Lymphoma	Neuroendocrine	Melanoma	random	all regions
Same-size CNN	0.547 / 0.546	0.564 / 0.605	0.774 / 0.788	0.599 / 0.623	0.656 / 0.669
same budget	0.625 / 0.623			0.682 / 0.700	0.701 / 0.719
YOLO11m	0.652 / 0.680	0.598 / 0.618	0.788 / 0.799	0.705 / 0.725	0.735 / 0.758
Frozen hibou-b	0.699 / 0.722	0.641 / 0.680	0.813 / 0.829	0.708 / 0.719	0.734 / 0.745
Karyo-T	0.675 / 0.698	0.619 / 0.653	0.809 / 0.831	0.722 / 0.745	0.745 / 0.760
Karyo-T×3	0.659 / 0.711	0.621 / 0.640	0.825 / 0.822	0.735 / 0.757	0.754 / 0.765

TTA changes the score scale, each model's threshold is re-chosen with TTA on its training-domain validation regions. The results without TTA remain the primary ones.

TTA raises held-out F1 for nearly every model and fold, by 0.01–0.04 (Table 7); the exceptions are both CNNs on lymphoma and the three-seed ensemble on melanoma, which do not change. Karyo-T gains +0.023, +0.034 and +0.022 held out (Holm-significant on lymphoma and neuroendocrine) and +0.009 to +0.022 on CMC. YOLO11m gains as much held out and the most on CMC (+0.017 to +0.028). The comparisons do not change: with TTA on both sides, no difference between Karyo-T and YOLO11m or hibou-b is significant on any fold or CMC setting (no Holm-adjusted p below 0.05; Table 10), and remains ahead of the CNN on held-out lymphoma (+0.152) and on CMC over all regions (+0.077 to +0.111). TTA does not fix the threshold transfer of Section 6.6: with TTA, the F1 lost to the validation threshold on lymphoma is 0.076 for Karyo-T (0.082 without), 0.078 for YOLO11m, 0.065 for hibou-b and 0.218 for the CNN, and on neuroendocrine 0.033–0.061. Its gain comes mostly from better ranking (held-out AP rises for every model, Karyo-T by 0.017–0.029). The one visible shift in the comparisons is that YOLO11m catches up on CMC over all regions (Karyo-T minus YOLO11m +0.000 to +0.003). TTA costs 8 forward passes: 13.5 s per region for Karyo-T on the M5, against 1.9 s without.

6.8 At similar size: three Karyo-T against YOLO11m (post hoc)

The ensemble of the three Karyo-T seeds has 18.1M parameters, close to YOLO11m's 20.05M, so comparing the two removes most of the size difference. We added this comparison after scoring and correct it as its own family of 18 (Appendix Table 11). The ensemble is ahead at all 18 point estimates: held out by +0.006, +0.023 and +0.037, on CMC random regions by +0.032, +0.021 and +0.037, and over all CMC regions by +0.011, +0.021 and +0.024, and still positive in all nine comparisons when both use TTA (+0.005 to +0.052). Four differences survive correction: CMC over all regions on the neuroendocrine and melanoma folds (+0.021 and +0.024), and, with TTA on both sides, held-out lymphoma (+0.032) and CMC random regions on the melanoma fold (+0.052). Two caveats limit this. The ensemble is one run per fold, so these intervals cover regions and slides but not training seeds; a seed-level comparison would need six more Karyo-T runs per fold. And YOLO11m was trained from scratch at under a third of Karyo-T's budget (Section 5). We read it as: at similar size and with these budgets, Karyo-T×3 is ahead of YOLO11m in four comparisons, not significantly different in the other fourteen, and ahead at every point estimate.

6.9 Adapting to the new tumour with a few labelled regions

If the main failure on lymphoma is the score scale, a little labelled data from the new tumour should recover much of it. We registered H6: a LoRA adapter trained on 1–5 labelled regions of the held-out tumour closes at least half of the gap to a Karyo-T that saw the tumour in training, averaged over folds whose gap is at least 0.02. We drew $k \in \{1, 2, 5, 10\}$ adaptation regions from the held-out tumour's training cases (never its test cases) and compared five options, each for 500 steps on windows drawn half from the k regions and half from the original training set, with τ re-chosen on the k regions: re-choosing τ only; training the heads only (449k parameters); LoRA [21] of rank 8 on the query and value projections (73.7k); LoRA on all four attention projections (147k); and full fine-tuning (6.05M).

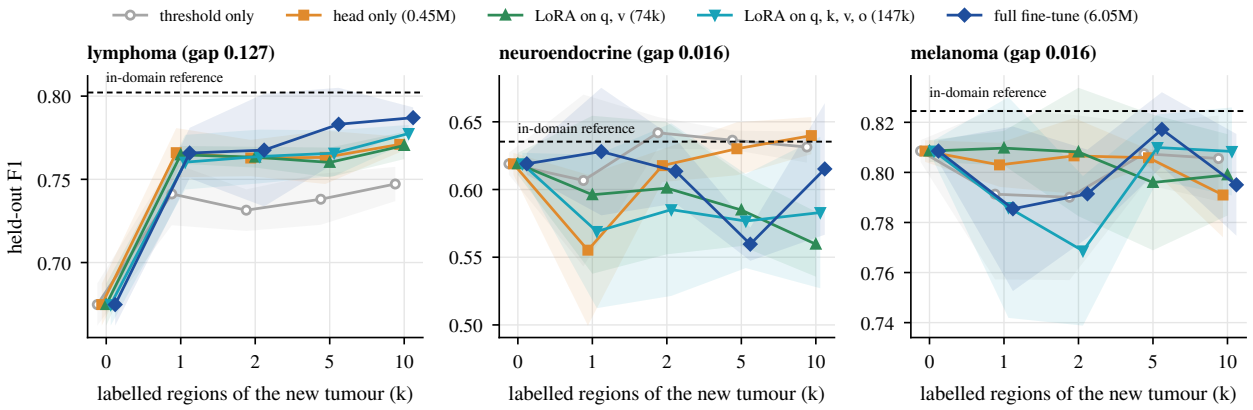

Figure 7. Held-out F1 after adapting Karyo-T with k labelled regions of the held-out tumour (mean over 3 seeds; bands ± sd). Dashed: Karyo-T trained with that tumour in its training set. Only lymphoma has a gap worth closing; there one labelled region recovers most of it. Where the gap is small, training on a few regions adds variance and can hurt.

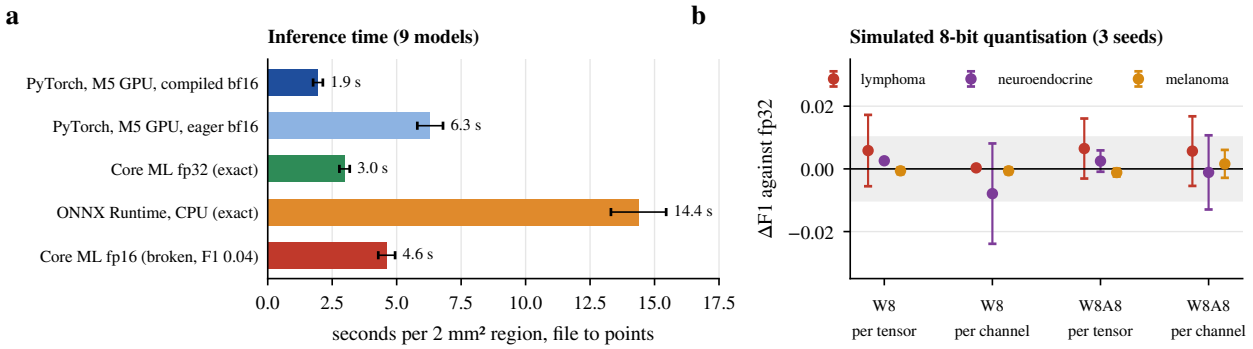

Figure 8. Deployment on one laptop (Apple M5), mean ± sd over the 9 models (3 folds × 3 seeds). (a) End-to-end time per 2 mm^2 region; "exact" means the same detections as PyTorch fp32. (b) Change in held-out F1 under simulated 8-bit quantisation of all 48 linear layers, with the threshold re-chosen on validation; the grey band is ±0.01.

Only lymphoma has a real gap (in-domain reference 0.802 vs 0.675; neuroendocrine and melanoma 0.016 each), so the registered rule decides H6 on lymphoma alone, where it is supported: LoRA at $k = 5$ reaches 0.760 ± 0.009 and closes 0.67 of the gap (Figure 7). One labelled region gives most of it (head 0.766, LoRA 0.765 at $k = 1$). Training beats re-choosing the threshold (+0.022 [+0.012, +0.030] at $k = 5$, Holm p 0.026) and lifts AP (0.833 vs 0.816), so it changes the scores as well as the operating point. Full fine-tuning is best (0.783 ± 0.022 at $k = 5$, 0.787 ± 0.006 at $k = 10$; +0.023 over LoRA, Holm p 0.018) and still 0.015 below the reference. LoRA on all four projections adds +0.006 [+0.001, +0.011] over query and value only, not significant after correction. Where the gap is small, adaptation adds noise or hurts: on neuroendocrine LoRA at $k = 5$ falls to 0.585 (−0.052 [−0.110, −0.002] against re-choosing τ, Holm p 1.00), with AP falling too.

6.10 Deployment

Speed. End to end (file to points) on the M5 GPU, Karyo-T takes 1.95 ± 0.19 s per 2 mm^2 region with `torch.compile` in bf16 (batch size 16–64 makes no difference), and 6.3 s uncompiled; reading and decoding are 13% of the time. bf16 and fp32 give identical detections at the operating threshold in all 9 models.

Tiling. ViT detections depend on where the window grid falls. Shifting the grid by 100 px leaves 0.51 of a standard ViT's detections in place (within 4 px), 0.81 of Karyo-T's architecture (before the training changes) and 0.93 of the CNN's. Running a 1,024 px crop in one pass instead of tiled agrees on only 0.63 of Karyo-T's detections at the threshold.

All reported numbers use the fixed tiling of Section 4.

8-bit quantisation (simulated, all 48 linear layers, 88% of the parameters; affine int8 [23] with ranges from 256 training-domain windows; τ re-chosen on validation) costs nothing measurable (Figure 8b): the mean F1 change is within ±0.01 on every fold for weights per tensor or per channel, with or without activations quantised, AP changes by at most 0.001, and detections at the fp32 threshold are identical for all 36 combinations of model and setting. Activation outliers are small (largest $|x|$ 7.3, in patch tokens of the last blocks; register tokens up to 4.9).

Export. ONNX (ONNX Runtime, CPU) and Core ML in fp32 reproduce PyTorch's detections exactly in all 9 models, at 14.4 and 3.0 s per region. Core ML in fp16 converts and runs but its outputs are wrong (held-out F1 0.04); we did not diagnose the cause.

6.11 Negative results: in-domain self-supervised pretraining

Two attempts to pretrain the encoder on MIDOG++ training-domain tiles before supervised training failed their first gate. LeJEPA [8] pretraining (class-token objective, weak colour augmentation) gave crop-classification fine-tuning no better than random initialisation, in-domain or held out, and its features were largely colour statistics (H1, not supported; one pretraining seed). A masked autoencoder [16] (mask 0.75, normalised-pixel targets, 10M tiles) produced patch features that located mitoses worse than a probe on six colour statistics (patch AP 0.089 vs 0.128) and cleared the random-initialisation floor but not the colour floor on a three-class centre probe (0.680 vs bar 0.714); we stopped before fine-tuning (H3, not supported at MAE defaults, one seed). In both cases the pretrained features carried little beyond colour.

7 Hypotheses

Table 8. Registered hypotheses and outcomes, after Holm correction where a bootstrap decides.

	Hypothesis	Outcome
H1	In-domain LeJEPA pretraining beats random initialisation on held-out F1.	Not supported (crop task, 1 pretraining seed).
H2	Conv stem, 2D RoPE and a multi-level neck close at least half of the held-out F1 gap between a standard ViT and the same-size CNN.	Supported on lymphoma; not established on melanoma (Holm p 0.076) or neuroendocrine.
H3	In-domain masked-image-modelling pretraining improves held-out F1.	Not supported (stopped at the probe gate, 1 seed).
H4	A ≤25M ViT distilled from frozen hibou-b comes within 0.02 F1 of it.	Not tested.
H5	Karyo-T beats RetinaNet's 7-fold mean (0.666), and is within 0.02 F1 of YOLO11m on MITOS_WSI_CMC.	Seven-fold part not tested (3 folds run; above RetinaNet on each); CMC part met on all three folds (+0.006 to +0.016 on all regions, +0.003 to +0.031 on random regions).
H6	LoRA on 1–5 labelled regions closes at least half of the in-domain gap.	Supported on the one fold with a gap (lymphoma, 0.67).

Table 8 summarises the hypotheses, all written before their experiments. H5, the headline, was registered over all seven MIDOG++ tumour types; we ran three, so its main part is untested, and the three folds we ran were also the folds on which Karyo-T was developed (only lymphoma was used to decide; the other two were run with frozen configurations, but their results from earlier stages were known when H5 was last amended). The analyses of Sections 6.2, 6.5, 6.7 and 6.8, the false-alarm measures on LUNG-MITO, the oracle threshold for every model, the Holm correction and the equivalence test were added after the main results were known, and are marked post hoc where they appear; the LUNG-MITO rules were registered before any model was scored on it.

8 Limitations

- **Three of seven folds.** We did not train on the other four held-out tumour types (mast cell tumour, lung carcinoma, soft-tissue sarcoma, breast carcinoma); a seven-fold mean is not available. Each fold has only 11–12 test regions, so per-fold intervals are wide, and one neuroendocrine region holds 55% of that fold's figures (Section 6.5).
- **External tests.** CMC is one tumour type from one laboratory, and its regions were selected by us with fixed rules, before scoring. LUNG-MITO adds a second laboratory and tumour type, but only recall can be read on it.
- **Baselines at their own budgets.** YOLO11m was trained from scratch, not from the COCO weights Ultralytics uses by default, and saw under a third of Karyo-T's training windows; only the CNN got Karyo-T's full budget, on lymphoma. A COCO-initialised YOLO11m at a larger budget may do better. Frozen hibou-b shares Karyo-T's recipe exactly but was pretrained on data we do not control, which is why we report it as a reference. We did not test distillation from it (H4).
- **Distillation not isolated.** Karyo-T's teacher is the CNN it is compared with, and we have no 3-seed run without the teacher at Karyo-T's budget (Section 6.2).
- **No comparison with the MIDOG 2025 leaderboard is possible.** Its hidden test set (365 cases, 12 tumour types, hotspot, random and hard-negative regions) contains human melanoma and canine mammary carcinoma but neither canine lymphoma nor human neuroendocrine tumour; its participants could train on all of MIDOG++, so melanoma was in-domain for them, and 10 of 14 also trained on MITOS_WSI_CMC [7]. Our numbers are leave-one-domain-out and our CMC regions are test-only, so neither our held-out F1 nor our CMC F1 can be placed on that leaderboard.
- **No stain normalisation** for any model.
- **Simulated quantisation**: accuracy only; integer kernels and their speed were not measured.

9 Conclusion

A 6M-parameter ViT trained from scratch on 278 regions is competitive with detectors 3 and 14 times its size on mitotic-figure detection under domain shift. With a convolutional stem, rotary positions, registers and a multi-level neck, plus hard negatives, weight averaging and distillation from a same-size CNN, it shows no significant difference from a 20M YOLO11m trained from scratch or from an 86M frozen pathology foundation model on three held-out tumour types of MIDOG++ and on the random regions of an external whole-slide test of canine mammary carcinoma, and its recall on human lung adenocarcinoma from another laboratory lies between theirs. The intervals are too wide to call it equivalent within 0.03 F1. It is ahead of the same-size CNN on the most shifted tumour, but that lead lies in how its threshold transfers, and on the other folds part or all of it may come from the CNN as teacher. Test-time augmentation helps nearly every model and changes none of these comparisons. For every model we trained, the largest remaining error on a shifted tumour is a threshold that does not transfer; for Karyo-T, one labelled region of the new tumour repairs most of it. The model runs in under two seconds per region on a laptop and survives 8-bit quantisation unchanged. We have not tested whether this holds on the four MIDOG++ tumour types we did not run.

Reproducibility

All code, tests and run outputs are in the project repository. The evaluation protocol and every decision were written down before the runs they govern, and every number in this paper is copied from a results log in which each row names its run files; the tables of the revision are generated from one statistics file by a script. The scorer was checked against the MIDOG organisers' evaluation code on real detections (identical TP, FP and FN in all 71 validation regions). Training was on a single Apple M5 laptop (Table 1); test-time augmentation for all 42 compared models took 69 h.

References

[1] M. Albahri et al. Recent advances in automated mitosis detection in digital pathology: A PRISMA-guided systematic review with evaluation-regime stratification (2018–2025). *Biomedicines*, 14(6):1369, 2026. doi: 10.3390/biomedicines14061369.

[2] J. Ammeling et al. Benchmarking foundation models for mitotic figure classification, 2025. arXiv:2508.04441.

[3] M. Aubreville, C. A. Bertram, T. A. Donovan, C. Marzahl, A. Maier, and R. Klopfleisch. A completely annotated whole slide image dataset of canine breast cancer to aid human breast cancer research. *Scientific Data*, 7:417, 2020.

[4] M. Aubreville, N. Stathonikos, C. A. Bertram, et al. Mitosis domain generalization in histopathology images – the MIDOG challenge. *Medical Image Analysis*, 84:102699, 2023.

[5] M. Aubreville, F. Wilm, N. Stathonikos, K. Breininger, T. A. Donovan, S. Jabari, M. Veta, J. Ganz, J. Ammeling, P. J. van Diest, R. Klopfleisch, and C. A. Bertram. A comprehensive multi-domain dataset for mitotic figure detection. *Scientific Data*, 10:484, 2023.

[6] M. Aubreville, N. Stathonikos, T. A. Donovan, et al. Domain generalization across tumor types, laboratories, and species – insights from the 2022 edition of the mitosis domain generalization challenge. *Medical Image Analysis*, 94:103155, 2024.

[7] M. Aubreville et al. Mitosis detection in the wild: Multi-tumor and context-aware generalization in the MIDOG 2025 challenge, 2026. arXiv:2606.07368.

[8] R. Balestriero and Y. LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics, 2025. arXiv:2511.08544.

[9] S. Banerjee, A. Teimoury, N. Porsche, et al. Beyond classification: Pathology foundation models as detection encoders for mitotic figures, 2026. arXiv:2607.28007.

[10] T. Darcet, M. Oquab, J. Mairal, and P. Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations*, 2024.

[11] C. de Souza. Metric discordance under MIDOG++ domain shift: a leave-one-domain-out audit of F1, emitted-detection confidence calibration, and ROI count error. *Journal of Pathology Informatics*, 2026.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[13] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1994.

[14] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu. UNETR: Transformers for 3D medical image segmentation. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[16] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[17] B. Heo, S. Park, D. Han, and S. Yun. Rotary position embedding for vision transformer. In *European Conference on Computer Vision*, 2024.

[18] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network, 2015. arXiv:1503.02531.

[19] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[20] F. Hörst, M. Rempe, L. Heine, C. Seibold, J. Keyl, G. Baldini, S. Ugurel, J. Siveke, B. Grüneisen, J. Egger, and J. Kleesiek. CellViT: Vision transformers for precise cell segmentation and classification. *Medical Image Analysis*, 94:103143, 2024.

[21] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[22] Z. Z. Iván, D. Hirling, G. István, C. Molnar, J. Ammeling, T. Micsik, K. Dobra, L. Kuthi, F. Sükösd, J. Fillinger, J. Moldvay, E. Tóth, M. Aubreville, V. Miczán, and P. Horvath. LUNG-MITO and stMIDOG++: A subphase-labeled mitotic dataset. Zenodo, 2026. Dataset, CC BY 4.0.

[23] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[24] G. Jocher and J. Qiu. Ultralytics YOLO11. https://github.com/ultralytics/ultralytics, 2024.

[25] H. Law and J. Deng. CornerNet: Detecting objects as paired keypoints. In *European Conference on Computer Vision*, 2018.

[26] Y. Li, H. Mao, R. Girshick, and K. He. Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision*, 2022.

[27] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision*, 2017.

[28] I. Loshchilov and F. Hutter. Decoupled weight decay regularization, 2019. arXiv:1711.05101.

[29] C. Marzahl and B. Napora. A bag of tricks for real-time mitotic figure detection, 2025. arXiv:2508.19804.

[30] D. Nechaev, A. Pchelnikov, and E. Ivanova. Hibou: A family of foundational vision transformers for pathology, 2024. arXiv:2406.05074.

[31] M. Oquab, T. Darcet, T. Moutakanni, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

[32] G. Percannella et al. Mitosis detection in domain shift scenarios: a Mamba-based approach, 2025. arXiv:2508.21033.

[33] A. Tarvainen and H. Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, 2017.

[34] D. Tellez, G. Litjens, P. Bándi, W. Bulten, J.-M. Bokhorst, F. Ciompi, and J. van der Laak. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical Image Analysis*, 58:101544, 2019.

[35] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, and R. Girshick. Early convolutions help transformers see better. In *Advances in Neural Information Processing Systems*, 2021.

[36] X. Zhou, D. Wang, and P. Krähenbühl. Objects as points, 2019. arXiv:1904.07850.

A Every comparison, Holm-corrected

Table 9 lists all 92 comparisons of the main family, Table 10 the 87 comparisons involving test-time augmentation, Table 11 the 18 comparisons of Karyo-T×3 with YOLO11m, and Table 12 the 39 equivalence tests.

Table 9. Every comparison in the paper, one family of 92: difference in micro F1 and, in brackets, the Holm-adjusted two-sided p (paired bootstrap, 10,000 draws; regions resampled on MIDOG++, slides on CMC). Bold: Holm-adjusted $p < 0.05$. The smallest attainable value is 0.0184. H6: options at $k = 5$ labelled regions (Section 6.9).

Setting	Comparison	Lymphoma	Neuroendocrine	Melanoma
held-out MIDOG++	Karyo-T − CNN	**+0.128 (0.018)**	+0.055 (1.00)	+0.034 (1.00)
	Karyo-T − CNN, same budget	**+0.050 (0.018)**		
	Karyo-T − YOLO11m	+0.023 (1.00)	+0.021 (1.00)	+0.021 (1.00)
	Karyo-T − frozen hibou-b	−0.024 (1.00)	−0.022 (1.00)	−0.005 (1.00)
	Karyo-T − Karyo-T, architecture only	+0.018 (1.00)	+0.058 (0.86)	**+0.066 (0.018)**
	Karyo-T×3 − Karyo-T	−0.016 (0.14)	+0.002 (1.00)	**+0.017 (0.018)**
CMC, random	Karyo-T − CNN	+0.115 (0.51)	+0.136 (0.89)	+0.116 (1.00)
	Karyo-T − CNN, same budget	+0.031 (1.00)		
	Karyo-T − YOLO11m	+0.015 (1.00)	+0.003 (1.00)	+0.031 (1.00)
	Karyo-T − frozen hibou-b	+0.004 (1.00)	+0.035 (1.00)	+0.002 (1.00)
	Karyo-T − Karyo-T, architecture only	**+0.102 (0.018)**	**+0.114 (0.018)**	**+0.100 (0.018)**
	Karyo-T×3 − Karyo-T	+0.018 (1.00)	+0.018 (1.00)	+0.006 (1.00)
CMC, all regions	Karyo-T − CNN	**+0.094 (0.018)**	**+0.090 (0.018)**	**+0.084 (0.018)**
	Karyo-T − CNN, same budget	+0.026 (1.00)		
	Karyo-T − YOLO11m	+0.006 (1.00)	+0.010 (1.00)	+0.016 (0.28)
	Karyo-T − frozen hibou-b	+0.004 (1.00)	+0.018 (1.00)	+0.012 (1.00)
	Karyo-T − Karyo-T, architecture only	**+0.070 (0.018)**	**+0.083 (0.018)**	**+0.075 (0.018)**
	Karyo-T×3 − Karyo-T	+0.006 (1.00)	+0.011 (0.51)	+0.008 (0.54)
CMC, hotspot	Karyo-T − CNN	**+0.089 (0.018)**	**+0.085 (0.018)**	**+0.077 (0.018)**
	Karyo-T − CNN, same budget	+0.017 (1.00)		
	Karyo-T − YOLO11m	−0.009 (1.00)	+0.007 (1.00)	+0.001 (1.00)
	Karyo-T − frozen hibou-b	−0.004 (1.00)	+0.014 (1.00)	+0.007 (1.00)
	Karyo-T − Karyo-T, architecture only	**+0.059 (0.018)**	**+0.079 (0.018)**	**+0.072 (0.018)**
	Karyo-T×3 − Karyo-T	−0.002 (1.00)	+0.007 (1.00)	+0.005 (1.00)
CMC, imposter-rich	Karyo-T − CNN	**+0.093 (0.018)**	**+0.073 (0.018)**	**+0.082 (0.018)**
	Karyo-T − CNN, same budget	+0.042 (0.14)		
	Karyo-T − YOLO11m	+0.030 (0.28)	+0.018 (1.00)	**+0.037 (0.038)**
	Karyo-T − frozen hibou-b	+0.020 (1.00)	+0.014 (1.00)	+0.024 (1.00)
	Karyo-T − Karyo-T, architecture only	**+0.073 (0.018)**	**+0.074 (0.018)**	**+0.066 (0.018)**
	Karyo-T×3 − Karyo-T	+0.014 (1.00)	+0.015 (1.00)	+0.015 (0.81)
H2	D = F1(Karyo-T, arch.) − mean(ViT, CNN)	**+0.141 (0.018)**	+0.093 (1.00)	+0.064 (0.076)
H6, $k = 5$	LoRA − re-chosen τ	**+0.022 (0.026)**	−0.052 (1.00)	−0.012 (1.00)
	full fine-tuning − LoRA	**+0.023 (0.018)**	−0.025 (1.00)	+0.021 (0.51)
	LoRA, all projections − LoRA	+0.006 (1.00)	−0.008 (1.00)	+0.014 (1.00)

Table 10. Post hoc. Test-time augmentation, its own Holm family of 87 (added after the main family was fixed). "TTA − plain": the same model with and without TTA; "both TTA": Karyo-T minus each model, TTA on both sides. Format as Table 9.

Setting	Comparison	Lymphoma	Neuroendocrine	Melanoma
held-out MIDOG++	CNN: TTA − plain	−0.001 (1.00)	**+0.041 (0.017)**	+0.013 (1.00)
	CNN, same budget: TTA − plain	−0.001 (1.00)		
	YOLO11m: TTA − plain	**+0.028 (0.017)**	+0.020 (1.00)	+0.011 (1.00)
	frozen hibou-b: TTA − plain	+0.024 (0.41)	+0.038 (1.00)	+0.016 (0.26)
	Karyo-T: TTA − plain	**+0.023 (0.017)**	**+0.034 (0.017)**	+0.022 (0.95)
	Karyo-T×3: TTA − plain	**+0.053 (0.017)**	+0.019 (1.00)	−0.003 (1.00)
	Karyo-T − CNN, both TTA	**+0.152 (0.017)**	+0.048 (1.00)	+0.043 (1.00)
	Karyo-T − CNN, same budget, both TTA	**+0.074 (0.017)**		
	Karyo-T − YOLO11m, both TTA	+0.018 (1.00)	+0.035 (1.00)	+0.031 (0.66)
	Karyo-T − frozen hibou-b, both TTA	−0.025 (1.00)	−0.027 (1.00)	+0.002 (1.00)
	Karyo-T×3 − Karyo-T, both TTA	+0.014 (0.27)	−0.013 (1.00)	−0.009 (0.19)
CMC, random	CNN: TTA − plain	+0.006 (1.00)	+0.031 (0.44)	+0.035 (0.25)
	CNN, same budget: TTA − plain	+0.018 (1.00)		
	YOLO11m: TTA − plain	**+0.035 (0.017)**	+0.008 (1.00)	+0.017 (1.00)
	frozen hibou-b: TTA − plain	+0.011 (1.00)	+0.017 (1.00)	+0.006 (1.00)
	Karyo-T: TTA − plain	+0.032 (0.28)	+0.020 (0.32)	+0.017 (1.00)
	Karyo-T×3: TTA − plain	+0.019 (1.00)	+0.012 (1.00)	+0.033 (1.00)
	Karyo-T − CNN, both TTA	+0.141 (0.38)	+0.125 (0.77)	+0.099 (1.00)
	Karyo-T − CNN, same budget, both TTA	+0.045 (1.00)		
	Karyo-T − YOLO11m, both TTA	+0.011 (1.00)	+0.015 (1.00)	+0.031 (1.00)
	Karyo-T − frozen hibou-b, both TTA	+0.025 (1.00)	+0.039 (1.00)	+0.013 (1.00)
	Karyo-T×3 − Karyo-T, both TTA	+0.005 (1.00)	+0.009 (1.00)	+0.022 (0.88)
CMC, all regions	CNN: TTA − plain	+0.005 (1.00)	**+0.017 (0.017)**	**+0.018 (0.028)**
	CNN, same budget: TTA − plain	**+0.018 (0.017)**		
	YOLO11m: TTA − plain	**+0.028 (0.017)**	**+0.017 (0.017)**	**+0.024 (0.017)**
	frozen hibou-b: TTA − plain	+0.011 (1.00)	+0.011 (0.79)	+0.010 (0.42)
	Karyo-T: TTA − plain	**+0.022 (0.017)**	+0.009 (0.57)	+0.011 (0.055)
	Karyo-T×3: TTA − plain	+0.021 (1.00)	+0.003 (1.00)	+0.009 (0.79)
	Karyo-T − CNN, both TTA	**+0.111 (0.017)**	**+0.083 (0.017)**	**+0.077 (0.017)**
	Karyo-T − CNN, same budget, both TTA	+0.031 (1.00)		
	Karyo-T − YOLO11m, both TTA	+0.000 (1.00)	+0.002 (1.00)	+0.003 (1.00)
	Karyo-T − frozen hibou-b, both TTA	+0.015 (1.00)	+0.016 (0.79)	+0.013 (1.00)
	Karyo-T×3 − Karyo-T, both TTA	+0.005 (1.00)	+0.005 (1.00)	+0.006 (0.19)

Table 11. Post hoc. Karyo-T×3 (18.1M, one ensemble of the three seeds per fold) minus YOLO11m (20.05M, 3 seeds), its own Holm family of 18 (added after scoring). The ensemble is one run per fold, so the intervals behind these p values cover regions (slides on CMC), not training seeds. Format as Table 9.

Setting	Comparison	Lymphoma	Neuroendocrine	Melanoma
held-out MIDOG++	Karyo-T×3 − YOLO11m	+0.006 (1.00)	+0.023 (1.00)	+0.037 (0.22)
	Karyo-T×3 − YOLO11m, both TTA	**+0.032 (0.004)**	+0.022 (1.00)	+0.023 (1.00)
CMC, random	Karyo-T×3 − YOLO11m	+0.032 (1.00)	+0.021 (1.00)	+0.037 (0.90)
	Karyo-T×3 − YOLO11m, both TTA	+0.016 (1.00)	+0.025 (1.00)	**+0.052 (0.036)**
CMC, all regions	Karyo-T×3 − YOLO11m	+0.011 (1.00)	**+0.021 (0.032)**	**+0.024 (0.004)**
	Karyo-T×3 − YOLO11m, both TTA	+0.005 (1.00)	+0.007 (1.00)	+0.009 (0.90)

Table 12. Post hoc. Equivalence of Karyo-T and each baseline within ±0.03 F1 (two one-sided tests on the same bootstrap draws as Table 9; margin set after the results were known), its own Holm family of 39: difference, 90% interval and Holm-adjusted TOST p. Bold would mark equivalence (Holm-adjusted $p < 0.05$); no comparison reaches it. An interval inside ± a reader's own margin shows equivalence at that margin before correction.

Setting	Comparison	Lymphoma	Neuroendocrine	Melanoma
held-out MIDOG++	Karyo-T − CNN	+0.128 [+0.096, +0.162] 1.00	+0.055 [+0.001, +0.088] 1.00	+0.034 [+0.005, +0.066] 1.00
	Karyo-T − CNN, same budget	+0.050 [+0.024, +0.081] 1.00		
	Karyo-T − YOLO11m	+0.023 [+0.002, +0.053] 1.00	+0.021 [−0.028, +0.096] 1.00	+0.021 [−0.004, +0.045] 1.00
	Karyo-T − frozen hibou-b	−0.024 [−0.045, −0.003] 1.00	−0.022 [−0.058, +0.027] 1.00	−0.005 [−0.034, +0.027] 1.00
	Karyo-T − Karyo-T, architecture only	+0.018 [−0.028, +0.059] 1.00	+0.058 [+0.020, +0.170] 1.00	+0.066 [+0.042, +0.117] 1.00
CMC, random	Karyo-T − CNN	+0.115 [+0.049, +0.176] 1.00	+0.136 [+0.037, +0.222] 1.00	+0.116 [+0.030, +0.196] 1.00
	Karyo-T − CNN, same budget	+0.031 [−0.021, +0.106] 1.00		
	Karyo-T − YOLO11m	+0.015 [−0.021, +0.050] 1.00	+0.003 [−0.033, +0.034] 1.00	+0.031 [−0.005, +0.083] 1.00
	Karyo-T − frozen hibou-b	+0.004 [−0.051, +0.056] 1.00	+0.035 [−0.022, +0.072] 1.00	+0.002 [−0.043, +0.035] 1.00
	Karyo-T − Karyo-T, architecture only	+0.102 [+0.060, +0.160] 1.00	+0.114 [+0.081, +0.145] 1.00	+0.100 [+0.070, +0.150] 1.00
CMC, all regions	Karyo-T − CNN	+0.094 [+0.065, +0.131] 1.00	+0.090 [+0.048, +0.149] 1.00	+0.084 [+0.046, +0.132] 1.00
	Karyo-T − CNN, same budget	+0.026 [+0.003, +0.062] 1.00		
	Karyo-T − YOLO11m	+0.006 [−0.006, +0.022] 0.46	+0.010 [+0.000, +0.023] 0.46	+0.016 [+0.006, +0.032] 1.00
	Karyo-T − frozen hibou-b	+0.004 [−0.011, +0.021] 0.33	+0.018 [+0.005, +0.030] 1.00	+0.012 [−0.001, +0.024] 0.33
	Karyo-T − Karyo-T, architecture only	+0.070 [+0.056, +0.090] 1.00	+0.083 [+0.066, +0.106] 1.00	+0.075 [+0.059, +0.097] 1.00