This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Supplementary Material
Can You Trust Frozen Hematology Foundation Models under Acquisition Shift?

Jai Kumar Sharma[1(✉)] and Peeyush Tapadiya[2]

[1] Virginia Tech
jaisharma@vt.edu
[2] Accenture
peeyush.tapadiya@accenture.com

This supplement provides supporting analyses for the main paper. Section labels A.1–A.16 match the citations in the main text. Unless noted, experiments follow the main-paper protocol: frozen embeddings, Acevedo as source, 5 source splits, and the 15-encoder benchmark. Mechanism probes in A.1–A.6 use the core 11-encoder set; A.7–A.16 use the full 15-encoder benchmark. The main paper is self-contained; this document provides additional controls, stress tests, and implementation details.

Main-paper claim	Supporting analysis	Section
Scanner shift is measurable	Scanner decodability, subspace, and projection analyses	A.1–A.3
Clean rank is unstable	Selection regret, paired bootstraps, source-target matrix	A.7, A.9, A.15
Exposure must be audited	Controlled exposure experiment	A.5
Class-specific failure modes localize the MLL23 collapse	Per-class confusion on MLL23	A.6
Calibration fails off-domain	ECE / adaptive-ECE / NLL / Brier and temperature scaling	A.13
Label-free selection fails under skew	Marginal entropy and agreement-style selectors	A.10–A.11
CBR mitigates skewed-prior TTA failure	CBR stress tests, edge cases, small batches, per-encoder harm counts	A.12
Resolution does not explain the rank flip	DinoBloom 224-vs-518 sensitivity	A.14
Probe-dependence is not a single-head artifact	Linear-head and local-geometry sensitivity	A.16

Sections A.1–A.6 analyze mechanisms and class-specific failures; A.7–A.11 analyze model selection; A.12–A.13 analyze adaptation and calibration; and A.14–A.16 provide resolution, source-target, and head-sensitivity controls.

A.1 Scanner identity is linearly decodable from frozen embeddings

To avoid a class-composition confound (domains differ in WBC prevalence, so a scanner classifier could exploit class frequency), we build a sample with **equal WBC-class composition across domains** (79 images per WBC class per domain, the largest common per-cell count), fit feature standardization on the **training split only**, and train (70/30 split, 5 balanced draws) a 4-way scanner classifier and a 5-way WBC-class classifier from each frozen encoder's features. We additionally decode the scanner **within a single WBC class at a time** (the tightest control) and report a permutation baseline (shuffled domain labels).

All 11 encoders (class-balanced)	scanner (chance 0.25)	WBC5 (chance 0.20)	within-class scanner
decodability	0.990–1.000 (mean 0.997)	0.782–0.971 (mean 0.891)	0.974–0.999 (mean 0.990)

Even with WBC-class composition equalized across domains and preprocessing fit only on training data, scanner identity is linearly decodable with accuracy 0.990–1.000 (mean 0.997), whereas WBC5 class decodability ranges 0.782–0.971 (mean 0.891); the permutation baseline sits at chance (mean 0.25). Crucially, scanner remains almost perfectly decodable *within a single WBC class* (mean 0.990), so this is not a class-prevalence artifact. In this auxiliary linear-decoding task acquisition is therefore separable at least as well as biology, indicating that acquisition information remains strongly represented in frozen blood-cell embeddings across model families. This extends the histopathology site-signature / batch-effect finding to single-cell hematology, across foundation-model families rather than a single pretraining recipe.

Scanner decodability is saturated across encoders and therefore does not by itself explain which encoder is most robust (Spearman ρ between scanner decodability and mean cross-dataset drop ≈ -0.09, not significant); the remaining variation appears associated with target-shift magnitude and encoder-specific class geometry (Sec. 5.1). A source-trained linear probe is fitted in a feature space that retains strong scanner-associated variation, so under a scanner-associated domain change the features shift along these directions relative to the source-fit boundary, contributing to the cross-domain macro-F1 collapse (Sec. 5). These results are consistent with a substantial first-moment component (a per-scanner mean offset) to the acquisition shift, which is what feature re-standardization corrects.

A.2 Class–scanner subspace analysis

For each encoder we measure the fraction of the WBC5 class-discriminative subspace energy that lies in the top-3 between-domain (scanner) subspace. Entanglement is small (0.03–0.16) and does not correlate with the cross-dataset drop (Spearman ρ=−0.03, p=0.94). Neither saturated scanner decodability (A.1) nor this global class–scanner overlap explains which encoder is most robust; the remaining variation appears more closely related to target-shift magnitude and encoder-specific class geometry (Sec. 5.1) than to a simple geometric feature-space property.

A.3 Scanner-direction projection as a diagnostic intervention

We project features onto the orthogonal complement of the top-3 (label-free) between-domain scanner directions, retrain the source linear probe, and re-evaluate cross-scanner. Projecting away these directions improves mean cross-target macro-F1 by +0.063 at zero in-domain cost (in-domain Δ = +0.000). This suggests that source-trained linear probes partly rely on scanner-associated directions. The effect is strongest for general encoders (DINOv2-S +0.145, DINOv2-L +0.126, CLIP-L +0.129) and weak for DinoBloom (+0.01–0.05) and BiomedCLIP (≈0), indicating that DinoBloom's MLL23 failure is not fully captured by a removable global scanner-mean axis but reflects its specific learned class geometry. Scanner-subspace projection is a diagnostic first-moment intervention, not the deployment method used in the main paper. It supports the low-order-shift interpretation and complements the CBR analysis.

A.4 Effective-robustness residuals are uninformative under saturated in-domain accuracy

We examined the confound via Taori effective-robustness residuals (above or below the in-domain-to-out-of-domain line). Because all frozen FMs have saturated, near-identical in-domain accuracy (0.98–0.997), this line is ill-conditioned (no spread on the x-axis) and the residual re-encodes cross-domain rank rather than a meaningful effective robustness above the line (DinoBloom-family residual ≈ −0.08 to −0.11 across targets). In this benchmark, the clean-vs-target rank matrix (Sec. 5.1, main-paper Fig. 1) is more informative than effective-robustness residuals.

A.5 Controlled exposure experiment

Setup. We fine-tune a clean encoder (DINOv2-S, no hematology pretraining) on one target scanner using WBC5 labels, re-extract features on all four domains,

train a fresh Acevedo probe, and measure the change in target macro-F1 relative to the unfine-tuned encoder.

Result. Fine-tuning on Matek-M8 leaves Acevedo in-domain unchanged (0.987→0.982, Δ−0.004) but raises the exposed scanner most (Matek 0.497→0.778, Δ+0.281) versus an unexposed scanner (MLL23 0.614→0.761, Δ+0.148); selective inflation = ΔMatek − ΔMLL23 = +0.133. Repeating for each scanner gives a diagonal-dominant matrix of macro-F1 gains over the unfine-tuned encoder:

fine-tune	test Matek	MLL23	Raabin
FT-Matek	+0.317	+0.189	+0.325
FT-MLL23	+0.278	+0.246	+0.198
FT-Raabin	+0.101	+0.061	+0.321

Fine-tuning produces a diagonal-dominant pattern overall: Matek and MLL23 gain most from same-scanner exposure, while Raabin is essentially tied after rounding (+0.325 under FT-Matek vs +0.321 under FT-Raabin). Subtracting the mean cross-scanner gain still leaves a positive selective component for all three targets: +0.128 for Matek, +0.121 for MLL23, and +0.059 for Raabin.

Interpretation. Scanner exposure can selectively inflate apparent robustness in a controlled supervised setting.

Limitation. This is supervised fine-tuning, not self-supervised pretraining, so it conflates seeing the scanner with learning the task (the clean signal is the selective diagonal component). It motivates exposure reporting and demonstrates the mechanism in principle, but does not prove DinoBloom-specific leakage inflation; DinoBloom's measured numbers do not show this inflation (Sec. 5.1).

A.6 Per-class confusion on MLL23

On the MLL23 target, DinoBloom-B mis-predicts lymphocytes as neutrophils (57% → neutrophil, only 16% correct; +26% → monocyte) and eosinophils as neutrophils (45%): minority WBC types collapse into the majority class off-scanner. DINOv2 retains lymphocytes (recall 0.82, 5-seed mean, matching the main paper) on the same images, its errors instead leaking eosinophils toward basophil rather than collapsing minority types into neutrophils. The Sec. 5.2 DinoBloom lymphocyte collapse is thus specifically a lymphocyte-to-neutrophil confusion, while DINOv2-B's robustness reflects a preserved class geometry rather than a global accuracy offset.

Having localized the failure to acquisition-sensitive class geometry, we next quantify its consequences for encoder selection.

A.7 Selection regret under balanced evaluation

For each target scanner, regret is the target macro-F1 of the oracle encoder minus the target macro-F1 of the encoder selected by in-domain accuracy.

Target shift from Acevedo	Linear regret	1-NN regret
Matek-M8 (smaller shift)	0.000	0.000
MLL23-Metafer (largest shift)	0.151	0.088
Raabin (mid shift)	0.064	0.129

Selecting by in-domain accuracy costs up to about 15 macro-F1 points on the most-shifted target (one would pick DinoBloom-L; RedDino is far better) and is near-free on the closest target. Clean accuracy is thus an unreliable deployment selector exactly where deployment risk is highest. This regret is computed on balanced targets; under realistic class imbalance the absolute numbers shift, and label-free alternatives to clean-accuracy selection also fail (A.10), so we recommend selecting encoders using representative cross-domain validation.

A.8 Generality across FM families

RedDino (MICCAI 2025; an RBC-focused DINOv2, described as RBC-focused rather than never trained on white cells) is the most robust encoder on the WBC MLL23 target (0.704, highest of 15), while the WBC-specialist DinoBloom-L falls to 10th of 15 (0.552) and the two pathology FMs diverge sharply (Lunit-DINO 0.609 vs Phikon 0.410). Specialization therefore does not guarantee robustness. This is a second-FM confirmation of the rank-stability issue, not an in-vs-out-of-pretraining contrast: RedDino, like DinoBloom, was trained on overlapping public hematology data.

A.9 Paired-bootstrap stability of the headline MLL23 rank differences

We assess whether the MLL23 rank differences are statistically stable. We compute a paired bootstrap of the macro-F1 difference for the four load-bearing comparisons: resample the 3,816 MLL23 test images with replacement (B=4,000), recompute each encoder's macro-F1 on the same resampled set (paired, removing shared example-level variance), average over the 5 source-split seeds, and take the difference. Point estimates reproduce the main-paper Table 1.

All four CIs exclude zero. The in-domain best, DinoBloom-L, is dethroned on MLL23: both RedDino and the smaller DinoBloom-S significantly outperform it. The pathology-FM divergence, with Lunit-DINO well above Phikon, is

Comparison (MLL23)	Δ macro-F1	95% CI	P(Δ>0)
RedDino − DinoBloom-L (in-domain #1)	+0.151	[+0.134, +0.168]	1.000
RedDino − DinoBloom-S	+0.033	[+0.017, +0.049]	1.000
DinoBloom-S − DinoBloom-L	+0.119	[+0.105, +0.133]	1.000
Lunit-DINO − Phikon	+0.200	[+0.184, +0.215]	1.000

also statistically stable. The paired design yields tighter, more appropriate intervals than per-encoder CI overlap because the encoders are scored on identical examples.

A.10 Marginal prediction entropy fails under realistic class-prior shift

Since clean accuracy mis-ranks encoders (Sec. 5.1), we assess whether a statistic from the source probe on unlabeled target images can rank them instead. We score five label-free signals (Spearman vs true target macro-F1 across encoders, per source-to-target): **conf** (mean max-softmax); **neg_ent** (−mean per-sample entropy); **neg_shift** (−RMS first-moment shift); **neg_predKL** (−KL of the predicted-label distribution from the source prior); and **marg_ent**, the entropy of the marginal prediction $H(\frac{1}{N}\sum_i p_i)$, a prediction-diversity / collapse detector. The figure in this section shows the balanced-evaluation result. The regret here uses the label-free selector benchmark: each unlabeled-target statistic ranks the 15 encoders and is scored against the true target macro-F1, averaged across the source-to-target evaluations. This differs from A.7, which reports Acevedo-source in-domain-accuracy selection regret on balanced targets.

Under balanced target sampling, marginal prediction entropy appears to select the oracle encoder:

predictor	mean ρ, MLL23 (balanced)	mean ρ, other targets	selector-benchmark regret (balanced)
clean accuracy	+0.27	+0.51	0.235
conf / neg_ent	+0.04 / +0.00	−0.16 / −0.23	0.287
neg_shift	+0.30	+0.59	0.236
neg_predKL	+0.50	+0.64	0.046
marg_ent	+0.89	+0.59	0.000 (= oracle)

This is a balanced-sampling artifact. Resampling the MLL23 target to realistic class priors:

target prior	marg_ent regret	clean-accuracy regret
balanced	0.002	0.162
peripheral-blood-like	0.290	0.151
neutrophil-heavy	0.372	0.139
lymphocyte-poor	0.252	0.122

Under every evaluated skewed prior, marg_ent's selection regret rises to 0.25–0.37 and clean accuracy beats it; the regret remains elevated across K=8–400, so it is not a sample-size effect. Under label shift a robust encoder should produce a skewed (low-entropy) marginal, so rewarding prediction diversity penalizes the right encoder. We therefore treat marginal entropy as a failure case, not a recommended selector.

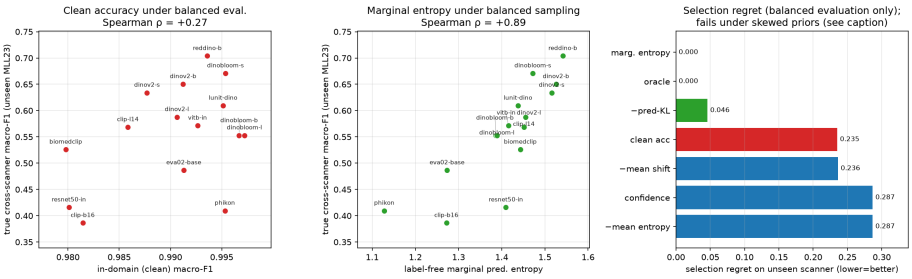


Fig. 1. (A.10) Marginal prediction entropy appears oracle-matching under *balanced* target sampling but fails under realistic skewed WBC priors. We report this as a failure case, not a recommended selection rule.

A.11 Comparison with agreement-style label-free selectors

Agreement-on-the-Line methods (formally cited in the main paper) predict out-of-distribution behavior from the agreement between models on unlabeled target data. We compare marginal prediction entropy with two agreement-style selectors: **consensus** (mean pairwise prediction-agreement of an encoder with all others on the unlabeled target) and **agl_acc** (an AgL-style accuracy estimate, clean accuracy scaled by the encoder's OOD/ID agreement ratio).

selector	selector-benchmark regret (balanced)	mean ρ vs true F1	needs
oracle	0.000	–	target labels
marg_ent	0.000	+0.89	one encoder, unlabeled target
agl_acc (AgL-style)	0.051	+0.88	all encoders' predictions + clean accuracy
consensus agreement	0.051	+0.88	all encoders' predictions
clean accuracy	0.235	+0.27	source labels only

Under balanced sampling, agreement-style selectors outperform clean accuracy. Because they were not evaluated under skewed priors here, their robustness to clinical WBC class-prior shift remains unestablished (they reward broad class usage and so may share the class-balance vulnerability of marginal entropy, but we do not measure this). Marginal prediction entropy, which was directly stress-tested under skew (A.10), fails consistently. We therefore do not identify a label-free selector with demonstrated robustness to the evaluated class-prior shifts; robust label-free encoder selection under label shift remains open.

The same class-prior assumption that destabilizes model selection also affects target-statistics adaptation, motivating the class-balanced analysis below.

A.12 CBR under label-shifted target priors

A.12.1 Main skewed-prior results

This is **one canonical experiment** that drives both Figure 3 and every CBR headline number: source = Acevedo, all 15 encoders, 5 source seeds, 3 targets, 6 target priors (balanced, Dirichlet-A, lymphocyte-poor, Dirichlet-B, peripheral-blood-like, neutrophil-heavy), 25 target draws per cell, fixed target batch N=400, primary metric fixed-WBC5 macro-F1. We measure macro-F1 gain over the no-adaptation baseline; the 18 scenarios are the 3 targets × 6 priors. CIs are a hierarchical bootstrap over scenarios (A.12.6).

method	mean [95% CI]	worst scenario	hurts
tgtstd	+0.032 [−0.006, +0.070]	−0.112	6/18
SHOT/IM	−0.029 [−0.087, +0.031]	−0.265	10/18
CBR	+0.059 [+0.046, +0.073]	+0.007	0/18
CBR-mean	+0.053 [+0.038, +0.065]	+0.002	0/18
CBR-oracle	+0.098 [+0.081, +0.113]	+0.023	0/18

Here tgtstd = global target feature standardization, CBR-mean = balanced mean only, and CBR-oracle = true-label-balanced statistics (a non-deployable upper bound). Across the 18 target-prior scenarios, CBR produced a positive mean gain in every evaluated scenario (mean +0.059, median +0.057, range [+0.007, +0.109], hurt 0/18), whereas global target standardization helps on balanced/mild priors but hurts in 6/18 (it falls to −0.07/−0.08 on the clinical and neutrophil-heavy priors), and SHOT/IM hurts in 10/18 (its diversity regularizer drives predictions toward a uniform marginal, which is wrong under skew). CBR uses pseudo-label-balanced, class-prior-reduced statistics: the balanced mean $\bar{\mu} = \frac{1}{|\mathcal{C}|}\sum_c \mathrm{mean}(X_t[\hat{c}{=}c])$ and a pooled within-class standard deviation. An ablation shows the balanced first-moment term drives the gain (+0.053 alone), tying CBR to the first-moment mechanism (A.1, A.3); the balanced within-class standard deviation adds little. CBR recovers about 61% of the true-label oracle, so pseudo-label noise costs something but CBR remains positive. CBR is label-free, training-free, and single-batch; it is transductive, pseudo-label-dependent, and evaluated on 5-class WBC. Decomposed per encoder (15 encoders × 18 scenarios = 270 cells, each cell a mean over 5 seeds × 25 draws), the mean CBR gain is positive for all 15 encoders, but 29/270 cells are negative: the two most skewed priors account for 23 of the 29 and the DinoBloom family for 15 (worst cell −0.090, DinoBloom-S, Matek/peripheral-blood-like; DinoBloom mean gains are smallest, +0.002 to +0.018, consistent

with their pretraining exposure to the targets, Sec. 5.1). Under the same decomposition tgtstd is negative in 96/270 cells (worst −0.24) and SHOT/IM in 149/270 (worst −0.37), each harming every encoder in at least one scenario. The headline "positive in all 18 scenarios" therefore holds at the scenario-mean level, not uniformly per encoder.

A.12.2 Pseudo-class edge cases

A pseudo-class predicted for no target cell in the batch is omitted from the balanced mean and standard deviation; a singleton pseudo-class contributes to the balanced mean but not the within-class variance (undefined for n=1); if the source probe collapses all target cells onto a single pseudo-class, CBR falls back to the source standard deviation. These rules keep CBR well-defined under the skewed and small batches that destabilize global target statistics, and are exercised by the small-batch stress test (A.12.5).

A.12.3 Target-prior definitions

Prior	Neu	Lym	Mon	Eos	Bas
Balanced	0.20	0.20	0.20	0.20	0.20
Peripheral-blood-like	0.62	0.30	0.05	0.02	0.01
Neutrophil-heavy	0.80	0.05	0.05	0.05	0.05
Lymphocyte-poor	0.40	0.02	0.19	0.19	0.20
Dirichlet-A	0.45	0.10	0.25	0.15	0.05
Dirichlet-B	0.10	0.50	0.15	0.20	0.05

Class order is neutrophil, lymphocyte, monocyte, eosinophil, basophil. The peripheral-blood-like prior approximates a normal adult WBC differential; the neutrophil-heavy prior stress-tests inflammatory or infection-like deployment batches (illustrative, not a clinical reference range). Each prior is applied by resampling the target test set to those class frequencies; the no-adaptation baseline uses the same resampled set, so the reported gain isolates the adaptation, not the resampling. Priors are applied by sampling without replacement (which can cap a rare class when the target pool is small); the realized class proportions for each draw are recorded in the canonical results file.

A.12.4 Comparison with BBSE label-shift correction

BBSE (Lipton et al., cited in the main paper) estimates target/source importance weights w_y by solving $\hat{C}w = \hat{\mu}$ ($\hat{C}$ = source confusion, $\hat{\mu}$ = target predicted marginal), then reweights posteriors. Over the same 18 canonical target-prior scenarios (A.12.1):

method	mean gain	worst	hurts
CBR	+0.059	+0.007	0/18
BBSE	−0.035	−0.051	18/18
CBR + BBSE	+0.055	+0.015	0/18

BBSE hurts macro-F1 in all 18 scenarios: it corrects the label prior (reweighting toward the estimated majority class), which can help balanced accuracy but hurts minority-class F1, and its prior estimate from a scanner-shifted, partly collapsing probe is unreliable. This indicates the failure is not a pure label-prior shift that prediction-space BBSE can solve; scanner-associated feature shift remains load-bearing. CBR+BBSE $\approx$ CBR, so label-shift correction is not the missing ingredient.

A.12.5 Small-batch and missing-pseudo-class stress test

The most demanding regime for CBR is tiny transductive target batches under realistic skew, where rare pseudo-classes are often absent. We sample K unlabeled target images at a skewed prior, adapt using those K (CBR vs vanilla tgtstd), and predict them; 15 encoders $\times$ 3 seeds $\times$ 20 draws.

K	prior	base F1	tgtstd Δ	CBR Δ	CBR hurt %	#pseudo (of 5)
16	PB-like	0.544	−0.003	+0.024	46.5	3.0
16	N-heavy	0.424	−0.011	+0.072	25.8	3.4
16	L-poor	0.470	+0.144	+0.092	19.9	3.8
32	PB-like	0.472	−0.017	+0.040	34.1	3.6
32	N-heavy	0.442	−0.027	+0.069	25.8	3.9
32	L-poor	0.435	+0.122	+0.085	15.3	4.3
64	PB-like	0.430	−0.057	+0.015	37.4	4.1
64	N-heavy	0.434	−0.065	+0.059	24.3	4.2
64	L-poor	0.422	+0.112	+0.081	12.8	4.5
128	PB-like	0.428	−0.062	+0.020	33.3	4.4
128	N-heavy	0.444	−0.078	+0.057	21.2	4.5
128	L-poor	0.434	+0.113	+0.082	8.7	4.7

PB-like = peripheral-blood-like; N-heavy = neutrophil-heavy; L-poor = lymphocyte-poor. Aggregated over the 12 K-prior cells, vanilla tgtstd has mean +0.014 (worst −0.078, hurts in 8 of 12 cells) while CBR has mean +0.058 (worst +0.015, positive in every cell): CBR is positive on average in every K-prior cell, even at K=16 with absent rare classes. The per-draw harm rate is non-trivial on the smallest batches under strong skew (up to about 46% at K=16) even though the mean is positive, so CBR is reliable in expectation but high-variance on tiny imbalanced batches, improving with K. We recommend K$\geq$32. In practice, these results support using CBR with batches of at least

roughly 32 unlabeled target cells, or aggregating statistics over multiple small batches when a deployment batch is extremely imbalanced.

A.12.6 Bootstrap interpretation

The CIs reported for CBR are a **hierarchical bootstrap clustered by scenario**: encoder×seed×draw observations within a target-prior scenario are not independent, so we first reduce each of the 18 scenarios to its mean gain, then resample the 18 scenario means with replacement (2,000 times), take the mean each time, and report the 2.5/97.5 percentiles. This respects the dependence structure rather than treating every encoder×seed×draw row as an independent replicate; the scenario range and hurt count are reported alongside the CI.

A.13 Calibration metrics and temperature-scaling transfer

Source-trained linear probe (source = Acevedo, 5 seeds), 15 encoders. Metrics: top-label ECE (15 equal-width bins), adaptive-ECE (15 equal-mass bins), NLL, and multiclass Brier. Temperature scaling fits T on the held-out source test split (minimizing NLL) and applies it to the target without refitting (the deployable case); oracle-T fits T on the target (using target labels, an upper bound, not deployable). The CBR arm re-standardizes target features before scoring.

metric	source (uncal.)	target (uncal.)	target + source-T	target + oracle-T	target + CBR	target + CBR + source-T
ECE	0.004	0.348	0.315	0.070	0.290	0.254
adaptive-ECE	0.003	0.348	0.315	0.073	0.290	0.254
NLL	0.031	3.200	2.515	1.205	2.213	1.772
Brier	0.013	0.783	0.750	0.597	0.683	0.652

Calibration collapses off-domain (ECE 0.004→0.348, about 80×; NLL 0.031→3.20; Brier 0.013→0.78); the reliability diagram (main-paper Fig. 2) shows target predictions below the diagonal (confidently wrong). Source-domain temperature scaling is insufficient for deployment on a shifted scanner: applying source-fitted T barely moves target ECE (0.348→0.315), because the source probe is already near-perfectly calibrated (fitted T≈1). Only oracle target temperature scaling substantially improves it (ECE→0.070, still above the in-domain 0.004: it improves but does not restore), and it requires target labels. CBR partially improves calibration (CBR alone 0.348→0.290 ECE; CBR+source-T 0.254 ECE, −27%; NLL −45%), but residual ECE (0.254) remains far above the in-domain 0.004. Robustness (Axis A) and calibration (Axis B) are distinct deployment failures that must both be evaluated per scanner/site.

Per-encoder, per-target ECE. Decomposed per encoder and target (15 × 3 cells), CBR+source-T lowers ECE in 43 of the 45 encoder×target cells; the two exceptions are Phikon and RedDino on Raabin (ECE increases by ≈0.03), and CBR alone helps in 40/45 (additionally failing for Lunit-DINO on MLL23/Raabin and ViT-B on Matek). Miscalibration is also uneven across targets: 13 of 15 encoders have their worst uncalibrated ECE on Raabin (0.31–0.63), so "partially improves calibration" holds for most, not all, encoder×target pairs, and residual miscalibration is largest exactly where calibration was worst.

Few-shot recalibration. Source-T does not transfer and oracle-T needs all target labels, but a small labeled per-scanner calibration set suffices. Fitting T on K labeled target images and applying it to the rest (15 encoders × 3 targets):

arm	target ECE
uncalibrated	0.351
+ source-T (no transfer)	0.308
+ few-shot target-T, K=4	0.165
+ few-shot target-T, K=8	0.124
+ few-shot target-T, K=16	0.100
+ few-shot target-T, K=32	0.088
+ few-shot target-T, K=64	0.081
+ oracle-T (all target labels)	0.074

About 16–32 labeled target images recover most of the calibration (within +0.02 of the oracle at K=32), turning Axis B into an actionable per-scanner protocol.

We finally test whether the conclusions depend on input resolution, source domain, or the operational head.

A.14 DinoBloom resolution sensitivity (224 vs 518)

The DinoBloom checkpoint name encodes a 224 input size, but its configuration inherits a 518 default from DINOv2, so running it at 224 (our benchmark's common resolution) could handicap it. On a controlled same-subset comparison (150 images per class, Acevedo source to MLL23 target), extracting each DinoBloom variant at both 224 and 518:

variant	MLL23 macro-F1 @224	@518	Δ(518−224)
DinoBloom-S	0.643	0.698	+0.054
DinoBloom-B	0.588	0.611	+0.023
DinoBloom-L	0.559	0.628	+0.068

Running at 518 helps DinoBloom (+0.02–0.07, most for the large model), so 224 mildly handicaps it. The qualitative conclusion is resolution-robust: at 518 the in-domain-best DinoBloom-L (0.628) is still beaten by the smaller DinoBloom-S (0.698). Giving every encoder its 518 input, DinoBloom-L (0.628) < RedDino (0.646) < DinoBloom-S (0.698), so the in-domain-best DinoBloom-L is still beaten by both a smaller DinoBloom and the RBC-focused RedDino. The exact rank and magnitude are resolution-dependent (DinoBloom-L's 10th-of-15 at 224 would improve to a middle rank at 518); we present the 224 benchmark as the common-resolution setting, while the reliability message (in-domain best is not most robust, and specialization and size do not guarantee robustness) holds at both resolutions.

A.15 Full source–target rank matrix (both probes)

Axis A's headline uses Acevedo as source. To check it is not an Acevedo artifact, we run the full source-target matrix (each domain as source) for **both probes**. We report Spearman ρ(in-domain rank, target rank) across the 15 encoders, and (for the linear probe) whether the in-domain-best encoder stays first on the target.

Linear probe.

source ↓ / target →	Matek	MLL23	Raabin	Acevedo
Acevedo	0.62 (no flip)	0.27 (flip)	0.43 (flip)	–
Matek	–	0.24 (flip)	0.52 (flip)	0.68 (no flip)
MLL23	0.46 (flip)	–	0.43 (no flip)	0.20 (flip)
Raabin	0.67 (no flip)	0.31 (flip)	–	0.53 (flip)

Across all 12 (source, target) cells, $\rho \in [0.20, 0.68]$, median 0.45 (mean 0.45), and the in-domain-best encoder is dethroned on the target in 8/12 cells. Clean linear-probe accuracy is a poor-to-moderate predictor of cross-dataset robustness from every source, and MLL23-as-target is the hardest (ρ 0.24–0.31) from every source, consistent with it being the largest distribution shift. When Acevedo is the target rather than the source, DinoBloom transfers best to it (rank 1), consistent with Acevedo being DinoBloom's held-out, canonical, high-SNR domain rather than evidence of leakage (Sec. 5.1).

1-NN probe. Repeating the full matrix with the 1-NN probe:

Across the 12 1-NN cells, $\rho \in [0.34, 0.88]$, median 0.65 (mean 0.67).

Probe-dependence. 1-NN rank transfer is higher on average than the linear probe (median 0.65 vs 0.45), so local nearest-neighbor geometry transfers more consistently than a source-fitted linear boundary. It is not universally predictive, however: Acevedo→Raabin drops to ρ=0.34 under 1-NN (below the corresponding linear ρ=0.43), and several 1-NN cells remain in the 0.55–0.65 range. Robustness rankings are therefore probe-dependent, and neither clean-domain probe is

source ↓ / target →	Matek	MLL23	Raabin	Acevedo
Acevedo	0.80	0.77	0.34	–
Matek	–	0.88	0.61	0.86
MLL23	0.77	–	0.65	0.66
Raabin	0.58	0.56	–	0.62

a universally reliable selector; MLL23-as-target remains among the hardest for both probes. The benchmark thus evaluates encoders together with the probe used to operationalize them, not a representation-only property.

A.16 The probe-dependence is not a single-head artifact

To verify that the linear mis-ranking is not an artifact of the specific linear head (logistic regression, C=1), we re-evaluate source→target rank transfer (Acevedo source, Spearman ρ across the 15 encoders) for a range of **source-only** heads (no target tuning): logistic regression at C ∈ {1e-4, 1e-2, 1, 1e2}, a linear SVM, a nearest-centroid classifier, a cosine nearest-centroid, and 1-NN. Every head uses the exact protocol of the main benchmark and A.15: the full Acevedo source, the same five shared source train/test splits, source-fitted standardization, and macro-F1 averaged over seeds. The logistic C=1 and 1-NN rows are therefore *by construction* the same experiment as the Acevedo row of A.15 and agree with it exactly.

head	in-domain F1	ρ→Matek	ρ→MLL23	ρ→Raabin	mean ρ
logistic C=1e-4	0.967	0.40	0.22	0.33	0.32
logistic C=1e-2	0.991	0.47	0.13	0.28	0.29
logistic C=1	0.990	0.62	0.27	0.43	0.44
logistic C=1e2	0.989	0.70	0.38	0.55	0.54
linear SVM	0.990	0.58	0.36	0.00	0.32
nearest-centroid	0.890	0.45	0.61	0.24	0.43
cosine centroid	0.892	0.58	0.69	0.09	0.45
1-NN	0.943	0.80	0.77	0.34	0.64

Every linear head reaches high in-domain macro-F1 (0.97–0.99) yet transfers weakly to the most-shifted target (MLL23 ρ 0.13–0.38), so the rank instability is not specific to the C=1 logistic probe: it persists across the full tested regularization range (C from 1e-4 to 1e2, six orders of magnitude) and under a linear SVM. The local-geometry heads (nearest-centroid, cosine, 1-NN) are more stable on MLL23 (ρ 0.61–0.77) at an in-domain cost (0.89–0.94). No head is universally reliable, however: on Raabin the linear SVM and cosine centroid fall to ρ 0.00/0.09, and even the best head on average (1-NN, mean ρ 0.64) is far from a safe selector. This confirms that clean-domain *linear* selection is unstable

across reasonable head choices on the most-shifted target, while local neighbor geometry transfers more consistently on average; robustness rankings depend on the operationalizing probe, not the representation alone.

Synthesis

Together, these controls support the main conclusion: favorable source accuracy and confidence do not guarantee target reliability, and the operational probe, the target class prior, and the acquisition domain all influence deployment performance.