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

Abstract. Frozen hematology foundation-model (FM) embeddings reach near-saturated in-domain white-blood-cell (WBC) accuracy, but clinical deployment demands reliability across scanners, sites, stains and preparation pipelines. We audit 15 frozen encoders (hematology, pathology, and general vision encoders) across four public single-cell acquisition domains along two axes: accuracy robustness and calibration. In-domain linear-probe macro-F1 is saturated (0.98–0.997), yet cross-dataset macro-F1 drops 34–72% and rankings re-order: DinoBloom-L, the in-domain best, falls to 10th of 15 on the most-shifted target (MLL23) at the benchmark's shared 224-px input, while RedDino and several general and pathology encoders outrank it. Rank transfer is *probe-dependent*: 1-NN retrieval is more stable on average than a source-fitted linear head (median ρ 0.65 vs 0.45), but neither clean-domain probe universally predicts target robustness. Calibration also collapses: source-trained probes are nearly calibrated in-domain (Expected Calibration Error [ECE] 0.004) but become confidently wrong off-domain (ECE 0.35), and source-fitted temperature scaling transfers poorly. We further audit pretraining exposure and identify MLL23 as corresponding to DinoBloom's internal cohort; because the only DinoBloom-held-out dataset is also our source domain, this benchmark cannot isolate exposure from scanner-associated distribution shift. Finally, label-free adaptation and marginal-entropy-based model selection appear safe under balanced evaluation but fail under realistic WBC class-prior shift. *Class-Balanced Re-standardization* (CBR), a training-free pseudo-label-balanced feature normalization, improves all evaluated target-prior scenario means and partially improves calibration, although encoder-level exceptions and residual miscalibration remain. These results argue that hematology FM benchmarks must jointly audit accuracy, calibration, exposure, and class-prior robustness.

Keywords: Foundation models · Domain shift · Robustness · Calibration · Hematology · Benchmark.

1 Introduction

Blood-smear classifiers are typically trained and benchmarked on one acquisition pipeline but deployed across scanners, laboratories, stains, and preparation protocols. For white-blood-cell (WBC) differential support this creates a dangerous failure mode: a model can be accurate on its development scanner yet systematically (and confidently) wrong elsewhere, where a second site cannot sanity-check it [3,6].

Hematology foundation models (FMs) are commonly compared using frozen linear-probe or nearest-neighbor accuracy on in-domain splits [1,2]. Such evaluation does not answer the questions that matter for deployment: whether clean rankings survive real acquisition-domain shift, whether predicted confidence remains calibrated off-domain, whether benchmark targets overlap the model's pretraining, or whether label-free test-time adaptation stays safe under the class imbalance of a real WBC differential.

We therefore *audit* frozen embeddings rather than train new classifiers. We train a source-domain linear probe (and a 1-NN probe), evaluate zero-shot across four public single-cell acquisition domains, and measure reliability along two axes: accuracy and calibration. We additionally audit dataset *exposure* and stress-test label-free adaptation under balanced and realistic target class priors.
Contributions. (1) A cross-domain *accuracy* audit of 15 frozen encoders (hematology FMs, pathology FMs, general SSL/VLM/ImageNet), showing that saturated in-domain linear-probe accuracy does not identify the robust encoder under scanner-associated shift, and that rank transfer is *probe-dependent* (1-NN ranks are more stable on average than a linear head, yet neither probe is a universally reliable clean-domain selector). (2) A *calibration* audit showing source-trained probes become confidently wrong off-domain and that source-fitted temperature scaling transfers poorly. (3) A *pretraining-exposure* audit showing public hematology benchmarks are exposure-ambiguous (we identify MLL23 as corresponding to DinoBloom's internal cohort), finding that these four datasets cannot isolate exposure from scanner-associated shift. (4) A label-shift stress test showing that standard label-free adaptation and marginal-entropy-based model selection, although effective under balanced evaluation, fail under realistic WBC priors. We evaluate **Class-Balanced Re-standardization (CBR)**, a training-free, frozen-feature instance of class-balanced normalization [33,34], which improves all evaluated target-prior scenario means and partially improves calibration.

2 Related work

Hematology FMs and WBC robustness. DinoBloom [1] and the RBC-specialized RedDino [2] are recent, widely used frozen hematology encoders; Tsutsui et al. [3] show supervised WBC CNNs degrade across imaging conditions, and WBCBench [7] benchmarks robust WBC classification under *class imbalance*. We instead audit *frozen* embeddings under *real* cross-dataset acquisition shift and stress-test class-prior robustness.

Scanner/site robustness in pathology FMs. ScanGen [6] and scanner-induced-shift work [27] establish scanner sensitivity for *tissue* FMs, attributing it to embedding/calibration drift rather than leakage. These motivate scanner/site reliability evaluation but do not study frozen single-cell hematology embeddings, calibration transfer, exposure ambiguity, or class-prior stress tests.

Calibration and exposure audits. Temperature scaling [36] and the off-distribution degradation of calibration [37] are established in general; we also report Expected Calibration Error (ECE), adaptive-ECE [38], NLL and Brier (Sec. 5), port these to frozen hematology FMs, and add the *transfer* test: does a source-fitted temperature survive a scanner change? PathBench [28] *prevents* pretraining-data leakage by strict eval/pretraining separation; we instead *audit* pretraining overlap and report it. Accuracy-on-the-Line [29] holds in-distribution, while Accuracy-on-the-*Wrong*-Line [30] shows it can break via label noise/nuisance features; we add a hematology cross-dataset case. (Effective robustness [4] and ImageNet-C [5] are the natural-image precedents.)

Label-free TTA and model selection. Our adaptation baseline is feature-space BN/AdaBN-style adaptation [8,9] (cf. CORAL [10], Tent [11], SHOT [12], and class-aware feature alignment [13]); *class-balanced / label-shift-robust normalization* [33,34] is an established sub-line we build on directly. We claim no new principle: CBR is its simplest frozen-FM, single-batch, training-free instance; unlike CAFA [13] it updates no weights, only re-estimating frozen statistics in a pseudo-label-balanced way. Our contributions are the first-moment diagnosis and the demonstration that vanilla BN-adaptation, SHOT/IM, and the label-shift method BBSE all silently fail under realistic class imbalance on hematology FMs, while the class-balanced instance does not. We also test a label-free encoder-*selection* statistic (marginal entropy), a *cautionary negative*, and compare with agreement-on-the-line performance prediction [31,32] (Sec. 5).

3 Method

Protocol. We freeze each encoder, extract its CLS/pooled features, fit source-domain standardization, and train an L_2-regularized logistic probe (and a 1-NN probe) on a source domain, then evaluate zero-shot transfer to held-out target acquisition domains. All encoders run at 224×224 (DinoBloom forced to 224, overriding the checkpoint's 518 default; resolution sensitivity in Suppl. A.14). We use the fixed five-class WBC intersection and evaluate the full source×target matrix, including a reverse-direction control.

Metrics. Our primary metric is macro-F1. We additionally report the Spearman rank correlation between in-domain and target encoder rankings, effective robustness, relative gap, per-class recall, and bootstrap 95% CIs over five source splits.

Label-shift evaluation. Because real blood differentials are class-imbalanced, we evaluate all label-free methods on the balanced target *and* on targets resampled to realistic/skewed class priors. The exact clinical and

neutrophil-heavy priors are listed in Suppl. A.12.3; they mimic peripheral-blood differentials and stress neutrophil-dominated deployment batches.

Test-time adaptation. Given a probe trained in source-standardized space, a standard label-free adaptation re-standardizes *target* embeddings with their own *unlabeled* batch statistics (`tgtstd`; feature BN-adaptation). Because the batch mean/std are dominated by the majority class, this breaks under label shift (Sec. 5). Our **CBR** estimates target statistics from *pseudo-label-balanced* class means $\bar{\mu}=\frac{1}{|\mathcal{C}|}\sum_c \mathrm{mean}(X_t[\hat{c}=c])$ and pooled within-class std $\bar{\sigma}$ (equal-weighting predicted classes *reduces* dependence on the target class prior, conditional on pseudo-label quality and class coverage; pseudo-labels $\hat{c}$ from the source probe), then transforms $(X_t - \bar{\mu})/\bar{\sigma}$, still label-free, training-free, single-batch. Empty pseudo-classes are omitted; singleton pseudo-classes contribute to the mean but not the pooled variance; if all target cells collapse to one pseudo-class we fall back to the source standard deviation.

4 Experimental setup

Source: Acevedo (PBC) [14] via BloodMNIST@224 [24] (CellaVision DM96), 10,298 WBC images. Targets (zero-shot): MLL23/Metafer [17], Matek-LMU/M8 [16], Raabin [15] (smartphone+Olympus). **Encoders** (15 frozen; families and full names in Table 1): DinoBloom [1], RedDino [2], Phikon [20], Lunit-DINO [21], DINOv2 [18], EVA-02 [19], BiomedCLIP [22], CLIP [23], ImageNet ViT-B [26] and ResNet-50 [25]; plus a supervised ResNet-18 baseline. **Pretraining-exposure audit:** DinoBloom reports training on all datasets *except* Acevedo; its internal "LabAnonymous" cohort is MLL23 [1,17] (same 41,906-image Munich Leukemia Laboratory dataset; the MLL23 descriptor cites DinoBloom as a prior user). So for DinoBloom, Acevedo (our source) is the only held-out dataset and Matek, Raabin and MLL23 are all in-pretraining, so there is no leakage-free *target*; we audit exposure rather than assume it. For the non-hematology encoders we found no documented *dedicated* exposure to these blood-cell datasets, though web-scale pretraining (e.g. CLIP) cannot be fully ruled out. We split at the image level (patient identifiers are not consistently available across these public datasets): the probe is trained on a 70% stratified split and the in-domain test is the held-out 30%, used only as a clean-ranking proxy; all deployment claims rest on cross-dataset transfer.

5 Results

Axis A: clean accuracy mis-ranks encoders. In-domain macro-F1 is saturated (0.98–0.997) yet cross-dataset macro-F1 drops 34–72% and re-orders: DinoBloom-L (in-domain #1) falls to 10/15 on MLL23 at the shared 224-px input (mid-rank at native 518, still below DinoBloom-S and RedDino; Suppl. A.14), beaten by RedDino (0.704) and even a pathology FM; the two pathology FMs diverge sharply (Lunit 0.609 vs Phikon 0.410), so specialization alone

Table 1. Cross-dataset leaderboard (linear-probe macro-F1, 5 seeds; source = Acevedo), sorted by MLL23. The "MLL23 #" column makes the re-ranking legible: the in-domain #1 (DinoBloom-L) falls to 10th of 15 on the most-shifted target (0.15 macro-F1 behind the best encoder, RedDino). Bold = strict column max. The supervised ResNet-18 (scratch, below the rule) is a non-frozen baseline, not part of the 15-encoder ranking. Paired-bootstrap 95% CIs for the key MLL23 comparisons (Suppl. A.9) all exclude zero. Exposure status audited in Sec. 4.

encoder	Acevedo[‡]	Matek	MLL23	MLL23 #	Raabin
RedDino *(hema)*	0.994	0.544	**0.704**	1	**0.450**
DinoBloom-S *(hema)*	0.995	0.613	0.671	2	0.341
DINOv2-B	0.991	0.595	0.650	3	0.444
DINOv2-S	0.988	0.505	0.634	4	0.248
Lunit-DINO *(path)*	0.995	0.409	0.609	5	0.288
DINOv2-L	0.991	0.555	0.588	6	0.315
ViT-B *(IN)*	0.993	0.616	0.571	7	0.439
CLIP-L/14	0.986	0.557	0.568	8	0.290
DinoBloom-B *(hema)*	**0.997**	0.635	0.553	9	0.448
DinoBloom-L *(hema)*[†]	**0.997**	**0.648**	0.552	10	0.385
BiomedCLIP	0.980	0.410	0.526	11	0.280
EVA-02 *(IN)*	0.991	0.486	0.486	12	0.378
ResNet-50 *(IN)*	0.980	0.357	0.416	13	0.304
Phikon *(path)*	0.995	0.448	0.410	14	0.265
CLIP-B/16	0.981	0.387	0.386	15	0.337
Sup. ResNet-18 (scratch)	*0.908*	*0.584*	*0.255*	–	*0.096*

hema=hematology FM, *path*=pathology FM, *IN*=ImageNet-supervised. [‡]Acevedo is the in-domain source; [†]in-domain #1 by unrounded macro-F1 (rounds to a tie with DinoBloom-B). "rank" = position on the most-shifted target MLL23 (1 = best of 15).

does not predict robustness. These gaps are statistically stable: paired bootstraps over the MLL23 examples put the RedDino−DinoBloom-L, DinoBloom-S−DinoBloom-L, and Lunit−Phikon differences at 95% CIs that exclude zero (Suppl. A.9). Spearman ρ(in-domain, cross-dataset rank) is as low as 0.27 on the most-shifted target (Fig. 1), so clean accuracy cannot discriminate robustness. This is not specific to the Acevedo source: across the full source×target matrix (Suppl. A.15), ρ stays low (median 0.45) and the in-domain-best encoder is dethroned in 8 of 12 source–target pairs. *Rank transfer is probe-dependent*: across the same 12 pairs 1-NN is more stable on average (median ρ 0.65 vs the linear 0.45; Suppl. A.15), so local retrieval geometry transfers more consistently, but 1-NN is not universal either (ρ 0.34 on Acevedo→Raabin) and is not a reliable clean-domain selector. This holds across logistic-C and linear-SVM heads (Suppl. A.16).

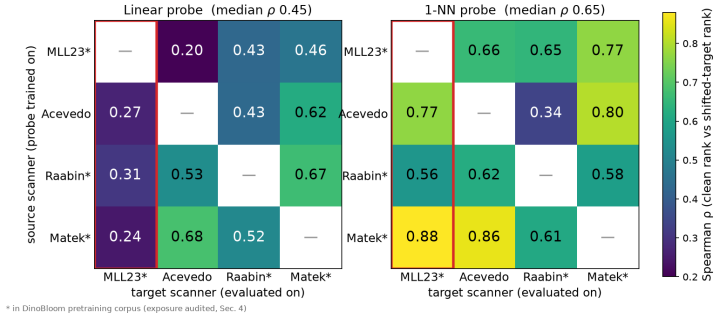

Fig. 1. Clean-to-target rank transfer is *probe-dependent*. Spearman ρ between in-domain and shifted-target encoder rankings (n=15), linear (left) and 1-NN (right). 1-NN is more stable on average (median ρ 0.65 vs 0.45) but not uniformly predictive (Acevedo→Raabin 0.34). MLL23 (red, largest shift) is hardest for both and is *not* leakage-free for DinoBloom (Sec. 4; $*$ = in DinoBloom pretraining corpus).

Exposure audit: no clear leakage-inflation signature. Pretraining exposure is an obvious confound, but the available datasets do not identify its effect cleanly. All three transfer targets were used in DinoBloom training (Sec. 4), so there is no leakage-free DinoBloom target: its target-side numbers measure transfer to *in-pretraining* domains, not clean held-out generalization. Conversely, in the reverse-source matrix (Suppl. A.15) DinoBloom transfers *best* to its held-out Acevedo and is not dominant on its training targets; this is inconsistent with a simple leakage-inflation explanation, but Acevedo is also an easier source/target domain, so exposure and domain difficulty remain confounded. A controlled fine-tuning experiment (Suppl. A.5) confirms scanner exposure can selectively improve same-domain transfer. We therefore report exposure status without claiming or excluding a DinoBloom-specific leakage effect.

Axis B: calibration collapses and source calibration does not transfer (Fig. 2). Source-trained probes are near-perfectly calibrated in-domain (ECE 0.004, NLL 0.03) but collapse off-domain (mean over 15 encoders × 3 targets: ECE 0.35, NLL 3.2; full metrics in Suppl. A.13). Crucially, **a temperature fitted on the held-out source does not transfer** to the target (ECE 0.35→0.32), since the source probe is already calibrated (T≈1); only *oracle* target temperature scaling *substantially improves* it (ECE→0.07), so calibration must be corrected per scanner. CBR alone improves target ECE 0.35→0.29, and adding the source-fitted temperature after CBR lowers it to 0.25 (43/45 encoder×target cells improve; Suppl. A.13), though substantial miscalibration remains (∼60× in-domain); CBR thus mitigates but does not solve calibration, and the two axes are distinct deployment problems.

Mechanism. Is the failure explained by synthetic stain/color perturbation, or by a measurable embedding shift? Synthetic stain corruptions do not reproduce the real failure (DinoBloom is stain-corruption-invariant yet collapses on Metafer); the dominant measurable effect is a shift in the per-feature means

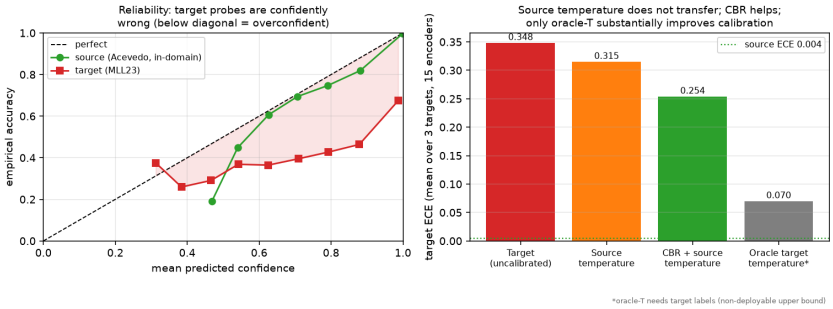

Fig. 2. Axis B (calibration). Left: reliability diagram (15 encoders pooled; Acevedo→MLL23): target probes sit below the diagonal (confidently wrong). Right: target ECE by arm (mean over 3 targets × 15 encoders): source temperature scaling barely helps; CBR + source-T partially helps; only oracle target-T (needs target labels, not deployable) substantially improves it, still above in-domain.

(the embedding's first moment). Per-class, DinoBloom-B/L collapse on lymphocytes off-distribution (recall 0.16–0.20, mis-called neutrophils) while DINOv2-B retains them (recall 0.82, 5-seed mean).

Label-free test-time adaptation: balanced evaluation hides failure under clinical priors. Because the measurable shift is low-order, we test whether label-free feature-statistics adaptation can mitigate it, in one canonical experiment (18 target×prior scenarios; Suppl. A.12). On balanced and mild priors global target standardization (`tgtstd`) helps, but on realistic neutrophil-dominated priors it *hurts* (clinical −0.07, neutrophil-heavy −0.08; hurts in 6/18 scenarios). Learned SHOT/IM is worse (mean −0.029, hurts 10/18), and a standard label-shift estimator, BBSE [35], hurts in *all* 18 (mean −0.035): it corrects the *label prior*, not the feature-space scanner shift. The failure is *not a pure label-prior shift solvable in prediction space*; a *class-balanced feature* statistic is better matched to it.

The class-prior problem also affects model selection: a label-free *selection* heuristic (marginal-prediction entropy) appears oracle-like under balanced target sampling but incurs 0.25–0.37 selection regret under skewed priors (Suppl. A.10–A.11): we identify no reliable label-free deployment selector.

For adaptation, however, class-balanced target statistics help: **CBR** (which equal-weights predicted classes to *reduce* dependence on the target class prior) is positive in all 18 evaluated target×prior scenarios (mean +0.059, range $[+0.007, +0.109]$, hurt 0/18; hierarchical bootstrap CI $[+0.046, +0.073]$, Suppl. A.12.6; Fig. 3). These are scenario means over 15 encoders: per encoder, mean gain is positive for all 15, though 29/270 encoder×scenario cells are negative (worst −0.09, mostly DinoBloom under extreme skew; vs 96/270 for `tgtstd`, 149/270 for IM; Suppl. A.12.1). An ablation shows the balanced first-moment term drives the gain (CBR-mean +0.053); CBR recovers 61% of the true-label

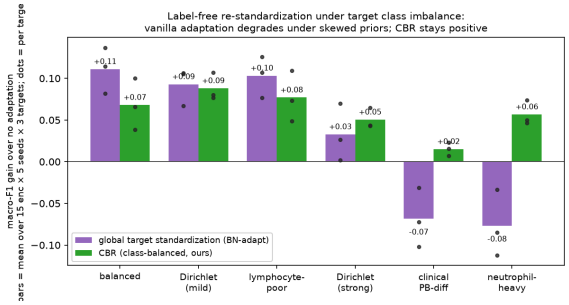

Fig. 3. Label-free re-standardization under target class imbalance, from one canonical experiment (15 encoders × 5 source seeds × 25 draws × 3 targets × 6 priors). Global target standardization (BN-adaptation) hurts on realistic neutrophil-dominated priors, while CBR (class-balanced, ours) is positive in all 18 *evaluated* target-prior scenario means (bars = mean gain over no adaptation, dots = per target; per-encoder breakdown: Suppl. A.12.1).

oracle and CBR+BBSE ≈ CBR. Small-batch stress tests (Suppl. A.12.5) show CBR stays positive on average even at K=16 but is noisy for tiny skewed batches; we recommend $K \geq 32$.

6 Discussion and limitations

Deployment implications. Select encoders using representative cross-domain validation rather than in-domain accuracy; disclose pretraining overlap; recalibrate confidence per scanner/site; and prefer class-balanced over global target-statistics adaptation when target batches are imbalanced. **Limitations.** The five-class WBC intersection; modest power for individual ρ over 15 encoders (we rely on CI-separated spread and rank re-ordering, and describe *scanner-associated cross-dataset* shift); no leakage-free DinoBloom target; probe-dependent selection failure (1-NN more stable on average but not universally reliable); transductive, pseudo-label-dependent CBR (∼61% of the oracle, $K \geq 32$, with per-encoder exceptions under extreme skew, Suppl. A.12.1); and one stain family per dataset.

7 Conclusion

In-domain linear-probe accuracy does not predict which frozen encoder is robust under scanner-associated shift; rank transfer is probe-dependent, both target accuracy and calibration degrade, and benchmarks are exposure-ambiguous. Class-balanced re-standardization gives a partial mitigation; we urge jointly auditing accuracy, calibration, exposure, and class-prior robustness for hematology FMs.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Koch, V., Wagner, S.J., Kazeminia, S., Sancar, E., Hehr, M., Schnabel, J., Peng, T., Marr, C.: DinoBloom: A Foundation Model for Generalizable Cell Embeddings in Hematology. MICCAI (2024). arXiv:2404.05022
2. Zedda, L., Loddo, A., Di Ruberto, C., Marr, C.: RedDino: A Foundation Model for Red Blood Cell Analysis. MICCAI (2025). arXiv:2508.08180
3. Tsutsui, S., Su, Z., Wen, B.: Benchmarking White Blood Cell Classification Under Domain Shift. ICASSP (2023). arXiv:2303.01777
4. Taori, R., Dave, A., Shankar, V., et al.: Measuring Robustness to Natural Distribution Shifts in Image Classification. NeurIPS (2020). arXiv:2007.00644
5. Hendrycks, D., Dietterich, T.: Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. ICLR (2019). arXiv:1903.12261
6. Carloni, G., Brattoli, B., Keum, S., Park, J., Lee, T., Ahn, C.H., Pereira, S.: Pathology Foundation Models are Scanner Sensitive: Benchmark and Mitigation with Contrastive ScanGen Loss. MedAGI Workshop, in Foundation Models for General Medical AI, pp. 44–53. Springer (2025). arXiv:2507.22092
7. Tian, X., Ma, X., Yang, T., Achim, A., Papiez, B.W., Watanaboonyongcharoen, P., Anantrasirichai, N.: WBCBench 2026: A Challenge for Robust White Blood Cell Classification Under Class Imbalance. ISBI (2026). arXiv:2604.10797
8. Li, Y., Wang, N., Shi, J., Liu, J., Hou, X.: Revisiting Batch Normalization for Practical Domain Adaptation. ICLR Workshop (2017). arXiv:1603.04779
9. Schneider, S., Rusak, E., Eck, L., et al.: Improving Robustness Against Common Corruptions by Covariate Shift Adaptation. NeurIPS (2020). arXiv:2006.16971
10. Sun, B., Saenko, K.: Deep CORAL: Correlation Alignment for Deep Domain Adaptation. ECCV Workshops, pp. 443–450 (2016). arXiv:1607.01719
11. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully Test-Time Adaptation by Entropy Minimization. ICLR (2021). arXiv:2006.10726
12. Liang, J., Hu, D., Feng, J.: Do We Really Need to Access the Source Data? Source Hypothesis Transfer for Unsupervised Domain Adaptation. ICML (2020). arXiv:2002.08546
13. Jung, S., Lee, J., Kim, N., Shaban, A., Boots, B., Choo, J.: CAFA: Class-Aware Feature Alignment for Test-Time Adaptation. ICCV (2023). arXiv:2206.00205
14. Acevedo, A., Merino, A., Alferez, S., Molina, A., Boldu, L., Rodellar, J.: A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. Data in Brief 30, 105474 (2020)
15. Kouzehkanan, Z.M., Saghari, S., Tavakoli, S., et al.: A large dataset of white blood cells containing cell locations and types, along with segmented nuclei and cytoplasm. Scientific Reports 12, 1123 (2022)
16. Matek, C., Schwarz, S., Spiekermann, K., Marr, C.: Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks. Nature Machine Intelligence 1, 538–544 (2019)
17. Shetab Boushehri, S., Kazeminia, S., Gruber, A., et al.: A large expert-annotated single-cell peripheral blood dataset for hematological disease diagnostics. Scientific Data 12, 1773 (2025). doi:10.1038/s41597-025-06223-x. Dataset: doi:10.5281/zenodo.14277609
18. Oquab, M., Darcet, T., Moutakanni, T., et al.: DINOv2: Learning Robust Visual Features without Supervision. TMLR (2024). arXiv:2304.07193
19. Fang, Y., Sun, Q., Wang, X., Huang, T., Wang, X., Cao, Y.: EVA-02: A Visual Representation for Neon Genesis. arXiv:2303.11331 (2023)

20. Filiot, A., Ghermi, R., Olivier, A., Jacob, P., Fidon, L., Mac Kain, A., Saillard, C., Schiratti, J.-B.: Scaling Self-Supervised Learning for Histopathology with Masked Image Modeling. medRxiv 2023.07.21.23292757 (2023)
21. Kang, M., Song, H., Park, S., Yoo, D., Pereira, S.: Benchmarking Self-Supervised Learning on Diverse Pathology Datasets. CVPR (2023). arXiv:2212.04690
22. Zhang, S., Xu, Y., Usuyama, N., et al.: BiomedCLIP: A Multimodal Biomedical Foundation Model Pretrained from Fifteen Million Scientific Image-Text Pairs. arXiv:2303.00915 (2023)
23. Radford, A., Kim, J.W., Hallacy, C., et al.: Learning Transferable Visual Models From Natural Language Supervision. ICML (2021). arXiv:2103.00020
24. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: MedMNIST v2 – A large-scale lightweight benchmark for 2D and 3D biomedical image classification. Scientific Data 10, 41 (2023)
25. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. CVPR (2016). arXiv:1512.03385
26. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR (2021). arXiv:2010.11929
27. Thiringer, E., Gustafsson, F.K., Ledesma Eriksson, K., Rantalainen, M.: Scanner-Induced Domain Shifts Undermine the Robustness of Pathology Foundation Models. arXiv:2601.04163 (2026)
28. Ma, J., Xu, Y., Zhou, F., et al.: PathBench: A comprehensive comparison benchmark for pathology foundation models towards precision oncology. arXiv:2505.20202 (2025)
29. Miller, J.P., Taori, R., Raghunathan, A., et al.: Accuracy on the Line: On the Strong Correlation Between Out-of-Distribution and In-Distribution Generalization. ICML (2021). arXiv:2107.04649
30. Sanyal, A., Hu, Y., Yu, Y., Ma, Y., Wang, Y., Scholkopf, B.: Accuracy on the Wrong Line: On the Pitfalls of Noisy Data for Out-of-Distribution Generalisation. arXiv:2406.19049 (2024)
31. Baek, C., Jiang, Y., Raghunathan, A., Kolter, J.Z.: Agreement-on-the-Line: Predicting the Performance of Neural Networks under Distribution Shift. NeurIPS (2022). arXiv:2206.13089
32. Saxena, R., Kim, T., Mehra, A., Baek, C., Kolter, J.Z., Raghunathan, A.: Predicting the Performance of Foundation Models via Agreement-on-the-Line. NeurIPS (2024). arXiv:2404.01542
33. Su, Y., Xu, X., Jia, K.: Towards Real-World Test-Time Adaptation: Tri-Net Self-Training with Balanced Normalization. AAAI 38(13), 15126–15135 (2024). arXiv:2309.14949
34. Vianna, P., Chaudhary, M., Mehrbod, P., et al.: Channel-Selective Normalization for Label-Shift Robust Test-Time Adaptation. CoLLAs 2024, PMLR 274, 514–533 (2025). arXiv:2402.04958
35. Lipton, Z.C., Wang, Y.-X., Smola, A.: Detecting and Correcting for Label Shift with Black Box Predictors. ICML (2018). arXiv:1802.03916
36. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On Calibration of Modern Neural Networks. ICML (2017). arXiv:1706.04599
37. Ovadia, Y., Fertig, E., Ren, J., et al.: Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. NeurIPS (2019). arXiv:1906.02530
38. Nixon, J., Dusenberry, M., Zhang, L., Jerfel, G., Tran, D.: Measuring Calibration in Deep Learning. CVPR Workshops (2019). arXiv:1904.01685