OUTNUMBEREDTALES: PREVALENCE-AWARE EVALUATION OF CONTRASTIVE LEARNING UNDER EXTREME BINARY IMBALANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Extreme class imbalance is a defining challenge across countless real-world domains where detecting rare events is critical. Despite this reality, the standard playbook for evaluating contrastive learning models relies heavily on artificially balanced test sets and symmetric accuracy metrics. Evaluating models in this artificial vacuum fails to reflect actual deployment conditions and obscures whether specialized, minority-aware algorithms provide any genuine benefit. Capturing a model's true efficacy demands a multifaceted approach. By examining performance across a broad spectrum of methods and decision-aware metrics, we demonstrate that different analytical lenses tell fundamentally different stories about model performance, exposing critical realities that conventional balanced testing actively conceals. Revisiting the problem through these varied lenses, we identify two distinct failure modes in existing minority-aware objectives. Geometrically, prototype-based contrastive losses degrade when fixed class prototypes conflict with the learned feature geometry. We introduce OrthoProto, an orthogonality hinge that resolves this conflict and surpasses the strongest published baseline wherever the failure is present, with gains scaling with the severity of the degradation. Statistically, batch-local objectives lose their supervisory signal when the expected number of minority samples per batch falls below one. We introduce CompMin, which maintains a census of minority representations and stays effectively invariant to batch size in this starved regime, a robustness we predict in advance from prevalence and batch size alone and then confirm. To ensure future research captures this multifaceted analysis, we release OutnumberedTales: a reproduction framework that implements published imbalance-aware objectives behind a common interface, pairs them with a decision-aware evaluation suite, and is designed for extension to new losses, datasets, and metrics.

1 INTRODUCTION

The imbalanced-learning literature primarily focuses on long-tailed multi-class problems involving hundreds of classes with smoothly decaying frequencies. These methods are typically evaluated by top-1 accuracy on a re-balanced test set (Zhang et al., 2023; Liu et al., 2019). Deployed classifiers often face a fundamentally different environment. Systems such as medical screening models, defect detectors, or rare-species monitors operate as binary classifiers where one class outnumbers the other at prevalences that can fall below 1%. Furthermore, the deployment base rate is uncertain and rarely equals the training rate. These two settings represent distinct challenges rather than the same problem at different scales. The simplex geometry of supervised contrastive learning degenerates to a single antipodal pair at $K = 2$ (Graf et al., 2021). Consequently, simplex-restoring (Zhu et al., 2022), prototype-contrast (Wang et al., 2021), and uniformity-target (Li et al., 2022) designs are exclusively analyzed at $K \geq 3$.

To the best of our knowledge, the first study to thoroughly analyze this behavior was conducted by Mildenberger et al. (2025). They demonstrated that supervised contrastive learning (SupCon) (Khosla et al., 2020) remains competitive under balance but collapses when the minority share falls between 5% and 1%. To address this collapse, they proposed two repairs. The first is SupMin, which applies class-level attraction exclusively to the minority. The second is SupProto, which enforces

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

attraction to two fixed antipodal prototypes. This paper characterizes a distinct failure mode for each repair family and proposes dedicated solutions. The first is a statistical vulnerability that we diagnose using a rule computable entirely from prevalence and batch size before training, which we confirm exactly where predicted. The second is a geometric issue traced to a measurable conflict between the fixed prototypes and the learned representation. Finally, we re-examine what "better" actually means once the evaluation protocol accurately reflects deployment conditions.

This re-examination forms our foundational evaluation argument. The field's default protocol reports one discrimination number at one artificial operating point, typically balanced accuracy on a balanced test set (Mildenberger et al., 2025; Zhu et al., 2022; Liu et al., 2019). We instead score every model through several lenses simultaneously. These approaches include prevalence-invariant ranking metrics, evaluating the same model at different prevalence ranges, and using a bounded decision-analytic score (Flores et al., 2025) that prices discrimination and calibration jointly over a declared prevalence band and error-cost ratio. Section 7 shows these lenses telling completely different stories about the same runs.

A geometric failure and OrthoProto. SupProto's prototypes are fixed before training, meaning the configuration they prescribe can conflict with the geometry the encoder actually learns. We quantify this mismatch with a *violation rate*: the fraction of minority embeddings that end pretraining within 90° of the majority centre, that is, samples still sitting on the majority's side of the representation space despite the prototypes prescribing the opposite. Under SupProto, the violation rate climbs from between 6% and 13% at a 5% minority share to between 80% and 100% at 0.5% prevalence. **OrthoProto** addresses this by adding a squared angular hinge to SupProto that targets the data's own majority direction instead of the frozen prototype, as shown in Section 4.1. This modification holds the violation rate between 3% and 6% across all tested settings, and its performance gain scales directly with the damage it repairs.

A statistical failure and CompMin. SupMin's minority supervision remains confined to the batch. Given a batch size B and prevalence π, a batch carries $p_B = 2B\pi$ minority views in expectation. When this expectation falls below one, the class-level term remains idle at most steps. Small contrastive batches are not an edge case. Consumer GPUs, memory-hungry multimodal or high-resolution encoders, and compute-poor labs all force their use, and contrastive methods are notoriously sensitive to exactly this parameter (Chen et al., 2020; He et al., 2020). **CompMin** removes the batch from the minority's supervision. Instead, a memory bank holds one momentum-averaged slot per minority training example so that every minority anchor is aligned against the entire class at every step as detailed in Section 4.2. Our empirical evaluation in Section 7.2.2 then confirms the performance prediction derived from π and B alone. Shrinking the batch into the starved regime costs SupMin up to 5 AUROC points while CompMin loses at most 1.5, a result that is significant in ten of twelve paired comparisons. At the standard batch size, CompMin matches SupMin to within a point everywhere. Its primary contribution is therefore robustness rather than raw accuracy.

OutnumberedTales. To ensure fair and extensible comparisons across the literature, we introduce OutnumberedTales. This reproduction framework is built to remove confounding degrees of freedom by implementing published imbalance-aware objectives behind a single interface and under one shared protocol. Every method is evaluated using one temperature, one encoder, and one probe per dataset family. This standardization ensures that performance differences reflect the underlying objectives rather than their implementation packaging. To comply with double-blind conference policies, the codebase is currently available anonymously at `https://anonymous.4open.science/r/outnumbered_tales-33C5/README.md`. The pretrained weights of every reported run will be fully released upon paper acceptance.

2 Related work

Contrastive learning and its geometry. InfoNCE-style objectives (van den Oord et al., 2018; Chen et al., 2020; He et al., 2020) contrast a positive pair against negatives drawn from the data marginal. SupCon (Khosla et al., 2020) extends this positive set to include all same-class samples present within the batch. While these methods offer theoretical guarantees under balanced conditions (Wang & Isola, 2020; Graf et al., 2021), they remain silent regarding extreme binary imbalance. Other approaches such as MoCo (He et al., 2020), XBM (Wang et al., 2020), and GSupCon (Hu et al., 2022) decouple the contrast set from the batch entirely by utilizing queues or per-example memories.

Long-tailed learning. Classifier-side corrections typically re-weight or re-margin the loss (Lin et al., 2017; Cui et al., 2019; Cao et al., 2019) or apply Bayes-consistent logit adjustment (Menon et al., 2021). Conversely, contrastive adaptations restore tail supervision through learnable centres (Cui et al., 2021), prototype injection (Zhu et al., 2022), capped positive sets (Kang et al., 2021), or uniform targets (Li et al., 2022). These methods are exclusively designed and evaluated at $K \geq 10$ using balanced test sets (Liu et al., 2019; Kang et al., 2020). Our baselines include logit adjustment and BCL as the classifier-side and contrastive representatives alongside in-batch adaptations of KCL and PaCo.

Binary imbalance. The direct predecessor to our work is the study by Mildenberger et al. (2025). To the best of our knowledge, they provided the first systematic investigation of supervised contrastive learning under binary imbalance. Their work introduced the SupMin and SupProto losses alongside a comprehensive diagnostics suite. We adopt and extend these diagnostics to ensure fair comparisons across all methods, but we depart significantly from their approach regarding evaluation metrics.

Evaluation. The tension between ROC and PR curves represents a long-standing challenge in classifier evaluation (Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015; Flach & Kull, 2015; Boyd et al., 2013). Recent studies highlight the fundamental limitations of both approaches. The value of the PR curve fluctuates with the class ratio while the ROC curve remains static (Richardson et al., 2024). Furthermore, optimizing AUPRC concentrates improvements on high-prevalence subpopulations (McDermott et al., 2024). Meanwhile, standard rank metrics like AUROC treat all errors equally and fail to account for the severe asymmetry between false positives and false negatives in real-world applications (Flores et al., 2025). To circumvent this blind spot, the decision-analytic literature explicitly assigns costs to different error types to better reflect actual deployment conditions (Elkan, 2001; Saerens et al., 2002; Vickers & Elkin, 2006; Gneiting & Raftery, 2007). Building upon this foundation, Flores et al. (2025) recently synthesized these asymmetric costs into a unified bounded decision score that evaluates model discrimination and calibration jointly.

3 SETTING

We consider binary classification with labels $y \in \{0, 1\}$ where class 1 represents the minority with prevalence $\pi = P(y = 1) \ll 1/2$. An encoder f and projection head g map an input x to $z = g(f(x))/\|g(f(x))\| \in \mathbb{S}^{d-1}$. Training follows the two-stage protocol of Mildenberger et al. (2025). This involves a contrastive stage on the imbalanced data with two augmented views per sample followed by a linear probe on the frozen encoder (Section 6). A batch of B samples yields $2B$ views. We denote $B_{\min}$ and B_{maj} as the minority and majority view sets respectively, τ as the temperature, and $p(i)$ as the augmentation partner of view i.

SupCon (Khosla et al., 2020) averages the InfoNCE loss for each anchor over all same-class views in the batch as positives against all other views as the contrast set. SupMin (Mildenberger et al., 2025) keeps the SupCon term for minority anchors and utilizes the NT-Xent loss for majority anchors (Chen et al., 2020). SupProto (Mildenberger et al., 2025) fixes two antipodal prototypes $p_{\min} = -p_{\text{maj}} \in \mathbb{S}^{d-1}$ that remain frozen for the entirety of training. The condition $\cos(z_i, p_{y_i}) \leq 0.5$ activates the prototype attraction only for samples still far from their target. Past that gate, NT-Xent alone governs a sample's position.

4 METHODS

4.1 ORTHOPROTO: SEPARATION MEASURED AGAINST THE LEARNED GEOMETRY

SupProto's gate relies on $p_{\min}$ which is a direction fixed before training and never updated. Consequently, the learned majority can inadvertently drift into the region the gate treats as safe. OrthoProto builds upon SupProto's existing machinery by introducing a term guided by the data's own majority direction rather than a frozen prototype.

$$\mathcal{L} \;=\; \mathcal{L}_{\text{SupProto}} \;+\; \lambda_{\text{sep}} \cdot \frac{1}{|B_{\min}|} \sum_{a \in B_{\min}} \big[\max(0,\; z_a \cdot \bar{z}_{\text{maj}})\big]^2, \tag{1}$$

where $\bar{z}_{\mathrm{maj}}$ is an exponential moving average (rate 0.9, detached, re-normalised) of the majority's batch-mean direction. The hinge penalises any minority embedding within 90° of the majority centre and is silent beyond orthogonality, so it does not compress the minority once separation is achieved. The 90° margin is not an arbitrary choice: the class-separability condition that Yi et al. (2025) derive for long-tailed representation geometry, $\frac{1}{2}\arccos(-1/(K-1))$, evaluates to exactly 90° at $K = 2$. Setting $\lambda_{\mathrm{sep}} = 0$ recovers SupProto exactly; we use $\lambda_{\mathrm{sep}} = 1$ in every reported cell. The state is a single d-vector and the added cost is $O(|B_{\min}|\, d)$. Figure 3 (Appendix A.2) illustrates the mechanism, and Table 8 (Appendix G) measures it: the fraction of minority embeddings inside the penalised hemisphere, 80–100% under SupProto in the extreme iNaturalist cells, stays at 3–6% under the hinge.

4.2 CompMin: the complete minority set as the positive set

CompMin maintains a memory bank $\tilde{Z} = (\tilde{z}_1, \ldots, \tilde{z}_m)$ containing one slot per minority training example. Each example owns the identical slot for the entire run. When a minority sample appears in a batch, the algorithm updates its slot as a momentum average of its detached and re-normalised embedding using $\mu = 0.5$. Both augmented views share this single slot.

Minority term. Every minority view $a \in B_{\min}$ aligns with every filled slot $j \neq \mathrm{slot}(a)$. Each alignment occurs in its own one-versus-majority contrast:

$$\mathcal{L}_{\min} = \sum_{a \in B_{\min}} \frac{1}{|J(a)|} \sum_{j \in J(a)} -\log \frac{\exp(z_a \cdot \tilde{z}_j/\tau)}{\exp(z_a \cdot \tilde{z}_j/\tau) + \sum_{n \in B_{\mathrm{maj}}} \exp(z_a \cdot z_n/\tau)}, \tag{2}$$

Here $J(a)$ represents the set of valid slots. The denominator holds only the single positive and the batch's majority views. Consequently, no other minority sample is ever repelled. Because the bank explicitly enumerates the entire class, the effective positive-set size remains strictly m rather than a random draw from $\mathrm{Binomial}(B, \pi)$. This ensures the quantity whose collapse breaks SupCon and thins SupMin remains constant by construction.

Majority term. Majority views receive instance-level contrast. The memory bank enters the denominator at the exact rate that minority views would naturally occur in a batch:

$$\mathcal{L}_{\mathrm{maj}} = \sum_{i \in B_{\mathrm{maj}}} -\log \frac{\exp(z_i \cdot z_{p(i)}/\tau)}{\sum_{j \in B_{\mathrm{maj}} \setminus \{i\}} \exp(z_i \cdot z_j/\tau) \;+\; w \sum_{j=1}^{m} \exp(z_i \cdot \tilde{z}_j/\tau)}, \qquad w = \frac{2B\pi}{\hat{m}}, \tag{3}$$

with $\hat{m}$ the number of filled slots. The weight w is an inclusion-probability correction: summing over all slots at weight w matches, in expectation, the contribution the in-batch minority views would have made, while removing its variance.

Aggregation. The final loss is $\mathcal{L} = (\mathcal{L}_{\min} + \mathcal{L}_{\mathrm{maj}})/2B$. Each anchor carries the fixed weight $1/2B$ instead of $1/|B_{\min}|$. This fixed weighting ensures the loss and gradient scale remain fully invariant to the realised minority count. The primary trade-off of the memory bank is staleness. A slot refreshes only when its corresponding example recurs, which happens roughly once per epoch. This specific trade-off helps us reason about the empirical patterns detailed in Section 7.2.2. CompMin provides a significant advantage when $p_B < 1$ but introduces a small cost where p_B is large. The per-batch computational cost is $O((2B)^2 d + 2Bmd)$ alongside a state memory requirement of $m \times d$ floats. Figure 2 and a comprehensive term-by-term breakdown are available in Appendix A.

5 Evaluation protocol

Mildenberger et al. (2025) evaluate using balanced accuracy on balanced test sets for iNaturalist (Van Horn et al., 2018) alongside AUROC at the native rate for medical datasets. This approach represents the field's default evaluation protocol (Liu et al., 2019; Cui et al., 2021; Zhu et al., 2022) and answers only a single question regarding how well the score ranks at one artificial operating point. A model trained at a low prevalence, such as 1% minority share, will ultimately be deployed at

some prevalence within a specific range, and sometimes with a defined tolerance for false positives, and possibly under strict probability constraints. Our protocol addresses these practical deployment questions directly.

Conventional lenses are retained. No single evaluation metric captures every dimension of model performance. We designed this protocol so that the strengths of one lens naturally compensate for the limitations of another. We retain AUROC in every table because its prevalence invariance is essential for deployment under uncertain base rates. This property ensures the metric remains meaningful regardless of the target environment while keeping our results directly comparable with prior work. However, AUROC does not retrieve us information about how a model behaves under a specific low-prevalence constraint. We also include balanced accuracy (BAcc) to maintain continuity with the field's standard evaluation practices despite it sharing this exact limitation.

Low prevalence metrics. We report both average precision (AUPRC) and the F1 score evaluated at the target deployment prevalence. We contrast these directly with their values at the conventional 50% balanced rate to highlight how they answer fundamentally different questions. The mathematical baseline for both AUPRC and F1 shifts directly alongside the class prevalence (Richardson et al., 2024; McDermott et al., 2024). We therefore strictly avoid comparing these metrics across varying prevalences or different datasets. Furthermore, AUPRC becomes markedly less reliable in this extreme regime. On identical training runs, its variance across seeds grows to a multiple of the ER-AUROC spread as the minority share falls. Figure 4 (Appendix B) illustrates this severe instability. To provide a robust alternative, we report **ER-AUROC** ($\mathrm{AUROC}_{0.1}$). This metric computes the standardised partial AUROC evaluated strictly over $\mathrm{FPR} \leq 0.1$ (Richardson et al., 2024). It successfully answers the top-of-ranking question that AUPRC is typically deployed to address while safely maintaining the inherent prevalence-invariance of the ROC curve.

Bounded decision score. AUROC and $\mathrm{AUROC}_{0.1}$ are rank statistics, invariant to any monotone rescaling, so neither can say whether a model's probabilities support a concrete decision at a declared cost (Flores et al., 2025; Hand, 2009). We therefore adopt the unmodified bounded decision log score S_{DCA} of Flores et al. (2025), which takes two declared inputs per dataset: a band $[a, b]$ of plausible deployment prevalences, and a cost ratio k of false positives accepted per additional true positive, fixing the break-even probability $c = 1/(1 + k)$ (Vickers & Elkin, 2006; Elkan, 2001). At each prevalence in the band the label-shift-transported, cost-optimal threshold (Saerens et al., 2002) is applied to the raw scores, the prior-adjusted cost-weighted accuracy is recorded (a perfect classifier scores 1), and S_{DCA} is the average over the band, uniform in log-odds; Flores et al. (2025) show this equals a clipped log score and so inherits proper-scoring-rule theory (Gneiting & Raftery, 2007) (computational details: Appendix C). Because S_{DCA} prices calibration alongside ranking, we follow them in splitting the two: $S_{\mathrm{DCA}}^{\mathrm{cal}}$, computed after isotonic regression, is the optimistic discrimination-only bound, so S_{DCA} is what the raw probabilities earn and $S_{\mathrm{DCA}}^{\mathrm{cal}}$ the ceiling of the ranking behind a well-fitted recalibrator.

6 Experimental setup

Datasets & Statistics. Three binary iNaturalist-2021 tasks (Van Horn et al., 2018), plants, insects and mammals, each at minority shares of 5%, 1% and 0.5% with the total training size fixed. Two MedMNIST tasks (Yang et al., 2023), BreastMNIST (native 27%, thinned to 20/15%) and PneumoniaMNIST (native 26%, thinned to 20/15/10/5/1/0.5%). Thinning subsamples the training minority only; validation and test are never thinned, and the full official test split is shared by all rungs of a task (Appendix E). Three seeds per cell (2000/2001/2002), identical across methods, five for BreastMNIST (2003/2004 added); mean $\pm$ sample sd. In tables, **bold** marks the best mean and underline the second best.

Methods. Two-stage contrastive: SupCon, SupMin, SupProto (Mildenberger et al., 2025), KCL (Kang et al., 2021), PaCo (Cui et al., 2021), and our CompMin and OrthoProto. Single-stage: BCL (Zhu et al., 2022) and logit adjustment (Menon et al., 2021). All contrastive losses share $\tau = 0.07$, ResNet-50 from scratch, projection dimension 128, batch 256 (32 in the batch study, learning rate scaled linearly), 350 pretrain epochs on iNaturalist and 100 on MedMNIST, full hyperparameters in Appendix D.

	native (27%)				20%				15%			
Method	AUROC	ER	S_{DCA}	$S^{\mathrm{cal}}_{\mathrm{DCA}}$	AUROC	ER	S_{DCA}	$S^{\mathrm{cal}}_{\mathrm{DCA}}$	AUROC	ER	S_{DCA}	$S^{\mathrm{cal}}_{\mathrm{DCA}}$
Logit adj.	72.9±3.6	56.9±4.8	71.3±8.9	80.1±1.9	67.1±5.0	54.3±1.6	69.9±3.5	78.6±2.1	62.8±4.5	52.1±2.3	**72.0±1.7**	77.7±1.7
BCL	71.4±5.6	55.5±5.2	72.6±2.7	79.8±1.8	63.6±6.8	52.9±2.8	72.8±2.1	77.6±2.5	64.0±3.8	53.4±1.4	70.4±4.6	77.5±1.6
SupCon	62.5±7.8	54.3±2.6	70.8±0.3	77.5±1.8	61.8±6.0	52.8±2.7	70.5±0.1	77.4±1.8	65.3±3.6	54.2±1.4	70.6±0.1	78.3±1.1
KCL	62.6±7.9	53.9±3.4	70.7±0.2	77.5±1.9	59.2±3.4	52.1±3.1	70.5±0.1	76.3±1.1	61.0±5.9	54.1±2.9	70.7±0.1	77.4±1.4
PaCo	70.3±7.5	57.3±4.7	72.3±1.6	79.6±2.6	61.5±4.8	53.5±1.5	70.7±0.2	77.5±1.3	64.5±5.1	53.4±3.0	70.7±0.1	77.7±1.3
SupMin	70.8±2.1	57.9±3.9	74.8±1.5	79.1±0.6	59.1±2.9	51.4±0.6	70.7±0.4	75.8±0.7	59.7±3.3	50.5±1.9	71.0±0.4	75.9±0.7
SupProto	79.3±2.5	66.0±3.6	74.0±2.4	82.3±1.2	67.3±6.4	56.5±2.9	70.8±0.3	78.4±1.9	63.7±4.2	54.6±2.0	70.8±0.3	77.7±1.5
CompMin	76.8±1.5	63.5±2.0	75.7±0.9	81.4±0.8	64.5±5.4	54.4±3.5	**73.2±1.8**	77.2±1.2	64.0±3.3	52.2±1.4	71.8±1.3	77.3±1.0
OrthoProto	**83.0±3.0**	**70.6±3.8**	**77.5±3.1**	**84.0±1.6**	**77.0±3.3**	**61.9±3.9**	72.7±1.5	**81.6±1.3**	**73.9±5.0**	**58.6±2.3**	71.8±1.6	**80.9±1.5**

Table 1: BreastMNIST, batch size 256, probe on the full balanced training subset. AUROC, ER-AUROC, and S_{DCA} raw and after isotonic recalibration ($S^{\mathrm{cal}}_{\mathrm{DCA}}$) at the clinical operating point of Table 5 ($\times 100$, mean$\pm$sd over $n{=}5$ matched seeds). **Bold**: best mean; underlined: second best.

	native (26%)				10%				1%				0.5%			
Method	AUROC	ER	S_{DCA}	$S^{\mathrm{cal}}_{\mathrm{DCA}}$	AUROC	ER	S_{DCA}	$S^{\mathrm{cal}}_{\mathrm{DCA}}$	AUROC	ER	S_{DCA}	$S^{\mathrm{cal}}_{\mathrm{DCA}}$	AUROC	ER	S_{DCA}	$S^{\mathrm{cal}}_{\mathrm{DCA}}$
Logit adj.	97.3±1.0	92.9±1.8	91.4±2.8	93.0±1.7	96.8±1.9	91.7±4.3	86.5±7.6	92.5±3.2	94.1±1.0	82.9±4.9	73.9±1.9	88.1±1.3	90.9±4.0	75.7±9.4	78.5±3.5	85.0±4.0
BCL	94.7±0.8	86.9±1.3	83.7±3.1	89.0±1.2	94.3±0.8	83.1±1.7	79.2±2.7	88.3±1.2	92.5±2.6	79.2±6.6	80.4±3.8	86.6±2.9	91.6±1.0	80.4±2.3	79.6±5.5	85.0±0.9
SupCon	99.4±0.1	97.3±0.4	93.5±3.1	96.4±0.2	98.1±0.1	94.7±0.4	90.1±4.8	94.3±0.3	88.1±2.6	68.3±4.2	72.7±5.5	82.0±2.6	84.1±1.4	63.3±2.1	69.1±1.8	78.3±1.1
KCL	99.4±0.1	97.5±0.2	96.1±0.2	96.6±0.1	98.3±0.3	95.4±0.3	93.9±0.7	94.5±0.3	90.0±1.3	72.3±1.2	73.9±4.7	83.3±1.4	85.5±1.6	64.5±4.2	67.3±12.5	79.8±1.2
PaCo	99.4±0.0	97.4±0.2	95.2±1.0	96.3±0.2	87.0±19.6	82.3±23.0	82.6±20.3	84.9±17.3	81.1±4.7	65.3±4.2	66.2±5.9	75.8±3.7	79.3±2.9	60.0±1.3	69.0±3.5	74.2±2.0
SupMin	99.5±0.1	97.8±0.2	**96.4±0.4**	96.8±0.4	98.3±0.0	95.0±0.3	93.6±0.3	94.3±0.2	92.9±0.8	81.6±3.6	83.4±0.3	86.5±1.8	90.1±3.1	75.1±7.2	80.3±2.3	83.9±3.3
SupProto	99.4±0.2	97.7±0.3	96.0±1.0	97.1±0.5	98.6±0.4	95.3±0.8	90.4±6.6	94.9±0.6	91.4±1.7	77.9±5.0	76.5±10.5	85.2±2.3	86.3±3.4	63.6±11.1	78.2±1.9	81.8±1.6
CompMin	**99.6±0.1**	**98.0±0.2**	95.9±1.6	**97.2±0.1**	**98.9±0.2**	**96.1±0.8**	91.3±5.7	**95.5±0.8**	92.4±1.8	80.9±2.9	82.0±2.3	86.1±2.3	90.1±2.0	75.3±4.7	80.9±1.0	84.1±1.5
OrthoProto	99.4±0.1	97.7±0.4	96.3±0.8	97.1±0.5	98.8±0.3	96.0±0.8	**94.2±0.7**	95.4±0.7	**96.0±0.5**	**91.1±1.7**	**90.2±0.5**	**91.4±1.3**	**95.0±1.4**	**88.8±2.9**	**87.5±1.5**	**89.8±2.1**

Table 2: PneumoniaMNIST, batch size 256, probe on the full balanced training subset. AUROC, ER-AUROC, and S_{DCA} raw and after isotonic recalibration ($S^{\mathrm{cal}}_{\mathrm{DCA}}$) at the clinical operating point of Table 5 ($\times 100$, mean$\pm$sd over $n{=}3$ matched seeds). **Bold**: best mean; underlined: second best.

Probe. The second stage trains a linear probe on the frozen encoder and its frozen batch-norm statistics, always on a class-balanced pool. On MedMNIST the pool is a balanced subset of the exact thinned training split used during pretraining, so the probe sees only images the contrastive stage saw. On iNaturalist we run two probes and report both boards, because the choice measures different things. The full-pool probe trains on the full balanced training set, whose labelled minority images were excluded from the imbalanced pretraining split: it scores the encoder's geometry with abundant labels, and is the configuration under which the published results of Mildenberger et al. (2025) are reproducible (Table 3, Appendix D). The thin probe trains on the balanced subset of the imbalanced split alone, so the pipeline uses no label beyond the pretraining budget (Table 4). Because single-stage baselines never access the full pool's extra minority labels, we report no single-stage scores on iNaturalist and compare them against contrastive methods only on MedMNIST, where budgets match. Comparisons among two-stage methods are unaffected, since within each board every contrastive method shares an identical probe.

7 Results

7.1 Medical panel

Tables 1 and 2 carry two column families that answer different deployment questions: the ranking columns (AUROC, ER-AUROC) measure whether the score orders minority above majority cases, prevalence-invariant and blind to probability quality; the decision columns (S_{DCA}, $S^{\mathrm{cal}}_{\mathrm{DCA}}$) measure what the raw probabilities earn once thresholded at the declared prevalence band and cost. We read the families separately because they disagree. All further metrics per cell are in Appendix H.

7.1.1 Ranking metrics

Saturated cells. Every two-stage method solves PneumoniaMNIST from native down to 10% (contrastive scores between 98.1 and 99.6 AUROC); we treat these cells as ties and draw no conclusions from them.

Cells with signal. The stable baseline finding: BCL sits significantly below SupMin on AUROC at every pneumonia rung from native down to 10% (-3.4 to -5.6) and at breast native (-2.6), and

logit adjustment behaves similarly; at mild imbalance (native 26–27% down to a 10% share), single-stage imbalance-aware training gives up several points to any two-stage contrastive method. Each of our methods reads against the loss it modifies. OrthoProto versus SupProto is the repair claim. It is significantly ahead on both AUROC and ER-AUROC in four of the five signal cells: $+9.7$ and $+5.4$ at breast 20%, $+10.2$ and $+4.1$ at 15%, $+4.7$ and $+13.2$ at pneumonia 1%, and $+8.7$ and $+25.3$ at 0.5%, with a positive but not significant $+3.7$ at breast native. The repair makes it the board leader: at breast 20% it leads all nine methods by 9.7 AUROC points ($+17.9$ AUROC and $+10.5$ ER-AUROC over SupMin, each significant), repeating at 15% with a lead of 8.6. CompMin versus SupMin is the robustness claim, and the medical board adds an accuracy note to it: CompMin sits significantly above SupMin at breast native ($+6.0$ AUROC) and 20% ($+5.4$) and tracks it within noise elsewhere. Breast seed spreads are large because training on 70 to 147 minority images is genuinely unstable. At pneumonia 5% the board compresses (OrthoProto first at 98.0, CompMin 0.3 below SupMin; Appendix H). The extreme rungs (1% and 0.5%) reverse the mild-imbalance pattern: at 1% OrthoProto is best ($+3.1$ over SupMin, significant) with logit adjustment and BCL (92.5) genuinely competitive, and at 0.5% (17 minority training images) it leads by $+4.9$ AUROC and $+13.8$ ER-AUROC over SupMin, both significant, while BCL and logit adjustment again sit above the SupMin family and SupCon collapses (88.1, 84.1).

7.1.2 Decision metrics at the clinical operating point

S_{DCA} answers the question the ranking columns cannot: thresholded at the declared cost, over the declared band, how good are the decisions the raw probabilities produce? Two reference points structure the reading: the best trivial policy (treat everyone or no one), which earns 70.4 at the breast operating point and 50.8 at the pneumonia one, and $S_{\mathrm{DCA}}^{\mathrm{cal}}$, the same score after isotonic recalibration, the ceiling the ranking could reach behind a well-fitted recalibrator, so the gap between the columns is the price of miscalibration alone.

BreastMNIST: the decision lens contradicts the ranking lens. At 20%, five of the seven contrastive methods score within half a point of the trivial policy (70.5 to 70.8). Their probabilities support no decision, and nothing in the ranking columns announces it. SupProto's AUROC at that cell (67.3) is nearly four points above BCL's (63.6), yet on S_{DCA} SupProto sits on the trivial policy's score (70.8) while BCL clears it (72.8). CompMin (73.2) and OrthoProto (72.7) are the only contrastive entries clearly above the policy, and logit adjustment's raw probabilities score below it (69.9). The $S_{\mathrm{DCA}}^{\mathrm{cal}}$ column locates the second failure mode where every breast model surrenders 4 to 9 points of S_{DCA} to miscalibration alone, making recalibration a requirement at these prevalences.

PneumoniaMNIST and family-specific deficits. The two-stage methods consistently produce usable probabilities at every prevalence rung. For single-stage baselines at mild imbalance levels, the performance deficit stems primarily from post-processing rather than representation. At 5% prevalence, logit adjustment initially trails OrthoProto by 21 raw points (72.6 against 93.6). However, recalibration closes 15.5 points of this gap (88.1 against 94.2; Table 28). At the extreme rungs, the decision lens separates the methods much more sharply than the ranking metrics. At 1% prevalence, the S_{DCA} for OrthoProto (90.2) sits 6.9 points above SupMin (83.4). This difference represents a substantial increase over the corresponding AUROC gap of 3.1 points, and this advantage firmly survives recalibration (91.4 against 86.5). At 0.5% prevalence, OrthoProto leads every method in both raw (87.5) and recalibrated (89.8) scores.

7.2 iNaturalist panel

7.2.1 Ranking metrics: SupProto's damage and OrthoProto's repair

Table 3 reports the iNaturalist board for the four minority-aware contrastive methods evaluated at the standard batch size. Among prior approaches, SupMin demonstrates the strongest performance while SupProto consistently degrades. SupProto falls significantly behind SupMin in five of the six extreme cells and reaches a deficit of -5.6 AUROC and -7.7 ER-AUROC on the insects 99 to 1 split. OrthoProto successfully corrects this specific degradation. It achieves the highest scores on insects at every prevalence and on every task at the 5% level. SupMin and CompMin retain the lead on plants at the 99 to 1 and 0.5% splits where the initial damage caused by SupProto remains relatively mild.

		Plants				Insects				Mammals			
Cell	Method	AUROC	ER-AUROC	AUPRC$_\pi$	BAcc	AUROC	ER-AUROC	AUPRC$_\pi$	BAcc	AUROC	ER-AUROC	AUPRC$_\pi$	BAcc
95:5	SupMin	96.3±0.3	87.3±0.9	68.8±0.8	89.9±0.5	92.6±0.3	82.7±0.9	66.5±2.2	84.1±1.1	86.9±1.0	74.5±1.2	50.8±4.9	78.2±1.0
	SupProto	96.0±0.5	87.9±1.1	71.4±4.3	89.5±0.8	92.2±0.3	82.8±1.7	67.9±4.0	83.2±0.3	86.7±0.8	73.7±0.3	45.6±2.6	77.9±1.4
	CompMin	96.4±0.2	88.5±0.6	71.3±0.3	90.6±0.5	92.3±0.5	82.5±0.9	67.0±0.5	84.2±0.6	86.7±0.3	74.0±0.4	48.3±4.6	78.3±0.6
	OrthoProto	**96.9±0.1**	**89.7±0.2**	**75.6±0.7**	**91.3±0.2**	**93.6±0.3**	**83.5±1.2**	**69.4±2.9**	**85.6±0.7**	**87.7±0.4**	**75.8±0.9**	**51.5±2.9**	**79.1±0.8**
99:1	SupMin	**94.0±0.3**	**81.8±0.8**	**30.3±3.4**	**86.8±0.3**	86.6±0.7	72.2±0.7	26.0±1.2	78.4±1.4	**83.4±0.1**	69.1±0.2	17.3±4.9	**75.6±0.5**
	SupProto	92.7±0.3	79.0±0.4	28.4±4.9	84.8±0.5	81.0±0.9	64.5±0.2	10.0±0.7	73.1±0.5	81.7±0.9	67.4±0.8	18.7±3.2	74.0±1.5
	CompMin	93.3±0.1	80.7±0.5	29.0±3.6	85.5±0.1	86.3±0.6	72.0±0.9	26.7±5.8	77.9±1.2	82.5±0.1	68.4±0.2	16.1±4.8	74.2±1.2
	OrthoProto	92.8±0.7	80.0±2.0	30.3±10.4	85.5±1.1	**88.6±1.1**	**75.7±0.6**	**35.7±1.8**	**79.8±1.9**	82.9±0.8	**69.3±1.2**	**19.4±3.3**	74.4±0.8
99.5:0.5	SupMin	93.0±0.5	**79.6±0.7**	14.2±0.6	85.4±0.7	84.0±0.5	68.4±1.0	18.5±6.4	75.4±0.6	**81.5±0.6**	67.2±1.3	**12.1±2.3**	**73.6±0.8**
	SupProto	91.1±0.3	73.2±0.9	10.5±2.7	83.9±0.4	80.8±0.3	63.6±0.9	5.9±1.3	72.8±0.8	80.3±0.7	65.7±0.6	12.1±0.9	72.0±1.1
	CompMin	**93.2±0.2**	79.5±1.4	17.5±1.2	**85.6±0.3**	83.5±0.2	67.8±0.9	15.0±5.8	75.4±0.2	81.0±0.7	**67.2±1.6**	11.7±2.4	73.0±0.5
	OrthoProto	92.2±0.5	78.6±1.1	**18.5±3.1**	84.6±0.7	**86.3±0.2**	**70.8±1.0**	**19.9±2.7**	**78.1±1.0**	81.3±2.0	66.9±2.1	9.9±2.9	73.4±1.5

Table 3: iNaturalist, batch size 256, probe on the full balanced pool. Metrics $\times 100$, mean±sd over seeds 2000–2002. **Bold**: best mean; underlined: second best.

		Plants				Insects				Mammals			
Cell	Method	AUROC	ER-AUROC	AUPRC$_\pi$	BAcc	AUROC	ER-AUROC	AUPRC$_\pi$	BAcc	AUROC	ER-AUROC	AUPRC$_\pi$	BAcc
95:5	SupMin	94.7±0.2	85.9±0.4	67.3±1.6	79.5±2.5	90.7±0.2	80.7±0.9	63.0±3.1	71.0±2.4	83.0±1.2	70.3±1.7	43.1±4.2	61.9±2.8
	SupProto	94.9±0.6	86.8±1.3	69.6±4.5	**83.5±1.3**	90.3±0.8	80.3±2.4	63.6±3.9	**74.6±2.0**	83.3±1.5	71.0±0.5	41.6±3.6	**67.9±1.9**
	CompMin	94.7±0.4	86.1±0.6	68.6±1.3	79.6±0.6	89.9±0.3	79.3±0.6	61.4±1.1	69.2±0.5	83.2±0.6	70.0±0.5	42.0±2.6	61.5±1.1
	OrthoProto	**95.3±0.8**	**88.3±0.8**	**74.4±2.4**	81.6±1.3	**91.5±0.4**	**81.4±0.8**	**64.6±0.7**	72.7±0.5	**84.6±0.7**	**71.2±1.2**	**45.7±3.5**	66.5±2.1
99:1	SupMin	87.8±1.3	74.5±1.0	**27.6±3.6**	61.5±3.5	80.2±1.6	66.9±1.9	17.8±2.7	55.2±3.0	71.9±2.6	58.2±2.3	9.1±4.4	52.2±1.7
	SupProto	87.8±1.5	71.0±2.3	18.1±8.1	**77.6±1.9**	71.7±1.0	57.3±1.9	4.1±1.5	**64.3±0.3**	**74.2±4.3**	59.5±3.6	7.2±2.9	**66.0±2.4**
	CompMin	**89.3±1.6**	**76.7±2.1**	26.9±7.0	65.7±1.5	80.8±1.7	66.1±1.3	15.0±4.9	55.7±0.8	73.3±1.6	**59.9±1.8**	**11.0±4.5**	53.0±1.4
	OrthoProto	85.3±1.0	72.9±1.3	23.8±7.1	63.6±2.3	**82.7±1.0**	**68.7±1.5**	**21.5±6.5**	58.2±2.7	69.4±2.8	58.4±1.7	7.8±2.3	52.2±0.9
99.5:0.5	SupMin	86.0±2.2	71.7±2.5	13.8±4.8	61.1±3.8	76.6±1.2	62.2±0.4	8.1±4.2	55.4±1.3	71.2±1.9	57.6±1.7	5.0±2.2	51.5±1.3
	SupProto	80.9±3.9	60.9±3.8	2.8±0.8	**73.8±4.1**	69.5±0.9	55.0±0.8	2.8±0.5	**61.4±2.3**	70.4±1.9	56.0±1.0	1.9±0.6	**64.7±1.7**
	CompMin	**87.1±1.2**	**72.5±0.8**	14.2±1.6	64.3±2.1	76.2±0.9	62.1±0.9	8.3±3.6	57.2±2.7	**72.4±1.0**	**58.4±0.9**	**5.5±0.3**	53.4±2.0
	OrthoProto	81.7±3.2	68.7±2.1	**15.1±1.3**	59.9±3.7	**78.3±1.6**	**62.8±2.4**	**10.0±4.7**	53.8±2.2	66.3±2.2	56.1±1.5	4.2±2.4	51.8±0.6

Table 4: iNaturalist, batch size 256, probe on the balanced subset of the thinned pretraining split (label-budget-matched). Metrics $\times 100$, mean±sd over seeds 2000–2002. **Bold**: best mean; underlined: second best.

The mechanism driving this behavior is measurable. The conflict measure introduced in Section 1 climbs under SupProto from 6 to 13% at 5% prevalence to 80 to 100% at 0.5% prevalence. SupProto's damage grows directly alongside this conflict. In contrast, OrthoProto holds this measure strictly at 3 to 6% everywhere. Its performance gain over SupProto scales predictably with the damage it repairs. The per-cell percentages and paired intervals are detailed in Tables 8 and 9 (Appendix G). The plants 99:1 cell resists the pattern, SupProto is significantly damaged (-1.3) yet the hinge recover is low ($+0.1$), at the mildest conflict reading (47%) among the extreme cells; we report it as the counterexample it is. The same logic holds against SupMin, OrthoProto surpasses it significantly exactly where SupProto's failure is severe (insects 0.5%, $+2.3$ AUROC; both insects cells on ER-AUROC) and not on plants, where the failure is mild and SupMin is already strong.

The repair under the matched budget. Table 4 rescores the same encoders with the thin probe, every score drops (1 to 15 AUROC points, most at the extreme rungs) and the repair claim splits by conflict severity. On insects, where the conflict is severe, OrthoProto stays significantly ahead of SupProto ($+11.0$ at 99:1, $+8.8$ at 0.5%) and above SupMin at 99:1 ($+2.5$); on plants and mammals at the extreme rungs, where it is mild, the advantage does not survive the matched budget and OrthoProto falls significantly behind SupMin (-2.4 to -4.9); CompMin posts the best mean in three of these six cells, though no lead over SupMin clears the paired bar (intervals: Table 11, Appendix G).

Why no decision metrics on iNaturalist. The score of Flores et al. (2025) takes the prevalence band and error cost as declared arguments, not free parameters. The medical datasets support such declarations and the iNaturalist tasks do not, so we report S_{DCA} on the medical panel only.

7.2.2 Batch-size robustness in the starved regime

CompMin is designed specifically for the small-batch regime. Because standard contrastive minority supervision relies entirely on local batch statistics, it vanishes under low prevalence and small batch sizes, constraints frequently imposed in practice by memory-intensive encoders. Extreme binary imbalance creates exactly the conditions that make our memory bank computationally feasible.

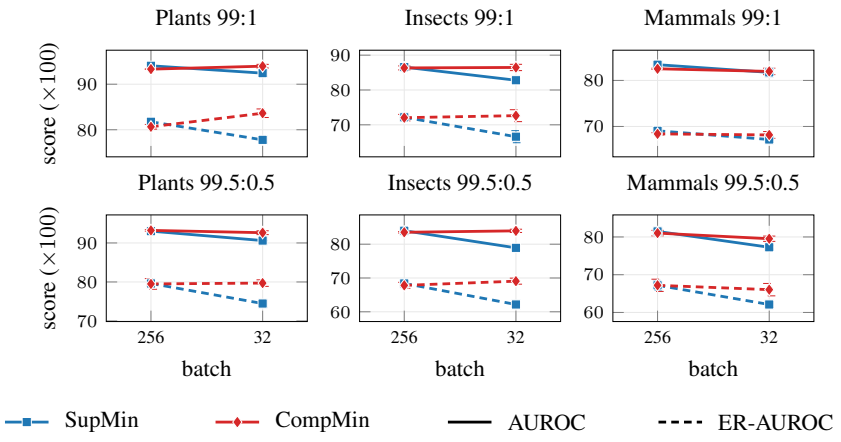

Figure 1: **Batch resistance.** Metrics at contrastive batch 256 and 32, one panel per starved iNaturalist cell. Paired intervals: Table 10 (Appendix G).

At standard batch sizes, this memory bank is on par with SupMin, providing minimal advantage. Table 3 confirms this behavior as the two methods perform within a single point of each other everywhere, with CompMin showing only two small significant deficits at the 99:1 split. The core capability of CompMin emerges when tracking the expected minority count $p_B = 2B\pi$. We hypothesize that SupMin's supervision degrades severely when $p_B < 1$, while CompMin's supervision remains completely intact. Figure 1 tests this directly. In the six starved cells where $p_B \leq 0.64$, shrinking the batch size costs SupMin up to 5.1 AUROC and 6.2 ER-AUROC points. Under these exact conditions, CompMin deviates by at most 1.5 points in either direction. Furthermore, the paired CompMin minus SupMin difference at batch 32 is strictly positive across all six cells on both metrics and statistically significant in ten of twelve comparisons (Table 10, Appendix G).

7.3 Representation diagnostics

Following Mildenberger et al. (2025) we compute their full diagnostic suite (SAD, SAA, CAD, CAC, GPU) plus our violation rate for every contrastive run; the record is reference material in Appendix F. The violation rate, the percentage of minority training embeddings that end pretraining within $90°$ of the majority centre, is the diagnostic the main argument uses. It climbs to 80–100% under SupProto in the extreme iNaturalist cells against 3–6% under OrthoProto's hinge, and jumps to 81–100% for SupMin at batch 32 in the starved cells (Table 8).

8 Conclusion

Under extreme binary imbalance, the choice of evaluation lens fundamentally alters what appears true. Metrics capturing ranking, prevalence-aware retrieval, and cost-explicit decision quality frequently disagree on the exact same predictions. By evaluating models across this spectrum, we isolate two distinct contrastive failure modes. The first is a geometric conflict, which OrthoProto's data-anchored hinge reliably repairs in direct proportion to the measured damage. The second is a statistical failure where batch-bound minority supervision collapses below one expected view per batch. In this starved regime, CompMin's memory bank successfully maintains representation quality precisely as predicted. Neither repair is a universal solution, and we document the conditions where each degrades. Finally, we release the OutnumberedTales framework so that future representation learning objectives can be evaluated through this same protocol.

Reproducibility statement

All experiments were conducted using the *OutnumberedTales* framework introduced above. The complete PyTorch codebase is provided alongside this manuscript to allow direct verification and

extension, and the pretrained weights for every reported run will be fully released upon paper acceptance to facilitate future evaluations without retraining. Every experiment is deterministically defined via a Hydra configuration, and the framework automatically archives the full configuration, per-epoch and final metrics, raw predictions, and representation diagnostics for each individual run. Finally, Appendices A through H exhaustively detail the exact loss algebra, metric implementations, decision-analytic operating points, training hyperparameters, dataset constructions, geometric diagnostics, and full per-cell results.

AI USE STATEMENT

During the preparation of this work, Claude Fable 5 and Claude Opus 5 (Anthropic) were used to assist with code debugging, implementing reproduction experiments, and verifying the correct computation of the representation diagnostics. Gemini (Google) was used to refine the manuscript's text for structural clarity and conciseness. Perplexity AI was employed during the literature review process to discover and explore related research. The authors thoroughly reviewed, edited, and validated all AI-assisted outputs and take full responsibility for the integrity, accuracy, and final content of this manuscript.

REFERENCES

Kendrick Boyd, Kevin H. Eng, and C. David Page. Area under the precision-recall curve: Point estimates and confidence intervals. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, volume 8190 of *Lecture Notes in Computer Science*, pp. 451–466. Springer, 2013. doi: 10.1007/978-3-642-40994-3_29.

Elizabeth S. Burnside, Jagpreet Chhatwal, and Oguzhan Alagoz. What is the optimal threshold at which to recommend breast biopsy? *PLOS ONE*, 7(11):e48820, 2012.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, 2019. arXiv:1906.07413.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*. PMLR, 2020. arXiv:2002.05709.

Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. arXiv:2107.12028.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. arXiv:1901.05555.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pp. 233–240, Pittsburgh, PA, USA, 2006. ACM.

Carl J. D'Orsi, Edward A. Sickles, Ellen B. Mendelson, and Elizabeth A. Morris. *ACR BI-RADS Atlas: Breast Imaging Reporting and Data System*. American College of Radiology, 5th edition, 2013.

Mark H Ebell, Isabella Locatelli, Yolanda Mueller, Nicolas Senn, and Kathryn Morgan. Diagnosis and treatment of community-acquired pneumonia in patients with acute cough: a quantitative study of decision thresholds in primary care. *British Journal of General Practice*, 68(676):e765–e774, 2018.

Charles Elkan. The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI)*, 2001.

Soudabeh Fazeli, James Stepenosky, Mary S. Guirguis, Beatriz Adrada, Rebecca Rakow-Penner, and Haydee Ojeda-Fournier. Understanding BI-RADS category 3. *RadioGraphics*, 45(1):e240169, 2025.

Peter A. Flach and Meelis Kull. Precision-recall-gain curves: PR analysis done right. In *Advances in Neural Information Processing Systems 28 (NIPS)*, 2015.

Gerardo A. Flores, Alyssa H. Smith, Julia A. Fukuyama, and Ashia C. Wilson. Aligning evaluation with clinical priorities: Calibration, label shift, and error costs. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, 2025.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Florian Graf, Christoph D. Hofer, Roland Kwitt, and Marc Niethammer. Dissecting supervised contrastive learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*. PMLR, 2021. arXiv:2102.08817.

David J. Hand. Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning*, 77(1):103–123, 2009. doi: 10.1007/s10994-009-5119-5.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. arXiv:1911.05722.

Zhijun Hu, Yong Xu, Jie Wen, Xianjing Cheng, Zaijun Zhang, Lilei Sun, and Yaowei Wang. Global-supervised contrastive loss and view-aware-based post-processing for vehicle re-identification. *arXiv preprint arXiv:2204.07943*, 2022.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations (ICLR)*, 2020.

Bingyi Kang, Yu Li, Sa Xie, Zehuan Yuan, and Jiashi Feng. Exploring balanced feature spaces for representation learning. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=OqtLIabPTit.

Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, Carolina C. S. Valentim, Huiying Liang, Sally L. Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, Vancouver, Canada, 2020.

Tianhong Li, Peng Cao, Yuan Yuan, Lijie Fan, Yuzhe Yang, Rogerio Feris, Piotr Indyk, and Dina Katabi. Targeted supervised contrastive learning for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Donna K. McClish. Analyzing a portion of the ROC curve. *Medical Decision Making*, 9(3):190–195, 1989.

Matthew B. McDermott, Haoran Zhang, Lasse Hyldig Hansen, Giovanni Angelotti, and Jack Gallifant. A closer look at AUROC and AUPRC under class imbalance. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024. arXiv:2401.06091.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations (ICLR)*, 2021.

David Mildenberger, Paul Hager, Daniel Rueckert, and Martin J. Menten. A tale of two classes: Adapting supervised contrastive learning to binary imbalanced datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10305–10314, 2025.

Mark I. Neuman, Michael C. Monuteaux, Kevin J. Scully, and Richard G. Bachur. Prediction of pneumonia in a pediatric emergency department. *Pediatrics*, 128(2):246–253, 2011.

Eve Richardson, Raphael Trevizani, Jason A. Greenbaum, Hannah Carter, Morten Nielsen, and Bjoern Peters. The receiver operating characteristic curve accurately assesses imbalanced datasets. *Patterns*, 5(6):100994, June 2024. doi: 10.1016/j.patter.2024.100994.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14(1):21–41, 2002. doi: 10.1162/089976602753284446.

Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):e0118432, 2015. doi: 10.1371/journal.pone.0118432.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Andrew J. Vickers and Elena B. Elkin. Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*, 26(6):565–574, 2006. doi: 10.1177/0272989X06295361.

Peng Wang, Kai Han, Xiu-Shen Wei, Lei Zhang, and Lei Wang. Contrastive learning based hybrid networks for long-tailed image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*. PMLR, 2020. arXiv:2005.10242.

Xun Wang, Haozhi Zhang, Weilin Huang, and Matthew R. Scott. Cross-batch memory for embedding learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2 – a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

Lingjie Yi, Jiachen Yao, Weimin Lyu, Haibin Ling, Raphael Douady, and Chao Chen. Geometry of long-tailed representation learning: Rebalancing features for skewed distributions. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10795–10816, 2023.

Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced contrastive learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. arXiv:2207.09052.

A TERM-BY-TERM BREAKDOWN OF THE PROPOSED LOSSES

This appendix restates the losses of Section 4, breaking down each symbol in the equations. Nothing here changes the objectives of Eqs. 2–1; the intent is that a reader can reimplement them from this section alone.

A.1 COMPMIN

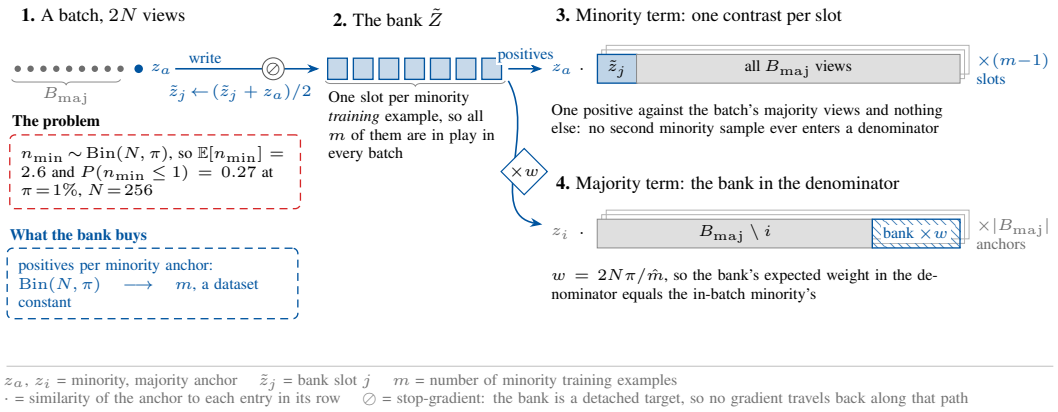

Figure 2: **CompMin.** The batch's minority supply collapses with the prevalence (**1**), so the bank (**2**) holds one slot per minority *training* example and is refreshed when that example recurs. It is then spent twice: as the positives of the minority term (**3**), and inside the majority term's denominator at the weight w that matches, in expectation, what the in-batch minority would have contributed (**4**).

The bank $\tilde{Z}$ and its update. The bank is m unit vectors plus a boolean *seen* mask ($J(a)$ = "the set of valid slots"). Every read is masked to the seen slots, so the random initialisation never enters a loss. A *write* is a plain copy of the (detached, re-normalised) embedding on first sight, and every later write is the momentum average $\tilde{z}_j \leftarrow \text{norm}\big(\mu\tilde{z}_j + (1-\mu)z\big)$, $\mu = 0.5$, so a slot retains roughly one step of history. The bank is refreshed with the current batch *before* either term is computed, so a minority sample drawn this step is already visible to this step's majority denominator. All bank entries are detached: they act as targets, never as parameters, so no gradient reaches a minority sample through a slot.

Slot assignment. Each minority training example owns one slot for the whole run: slot j is the rank of the example's dataset index among the minority indices. The bank is therefore the complete minority set by construction, one (momentum-averaged) entry per example, and a slot never blends different examples.

Minority term (Eq. 2), factor by factor.

(i) $\sum_{a \in B_{\min}}$ ranges over minority *views*, so a minority sample drawn once contributes two anchors.

(ii) $\frac{1}{|J(a)|}\sum_{j \in J(a)}$ averages over the anchor's positives, which are the filled bank slots other than the anchor's own.

Writing $s_j = z_a \cdot \tilde{z}_j/\tau$ and $s_n = z_a \cdot z_n/\tau$ for $n \in B_{\text{maj}}$, and p_j, q_n for the softmax weights of the positive and of majority view n within the pair's denominator, the gradient of one pair term is

$$\frac{\partial}{\partial z_a}\Big[-\log \frac{e^{s_j}}{e^{s_j} + \sum_n e^{s_n}}\Big] = -\frac{1}{\tau}\Big[(1-p_j)\,\tilde{z}_j - \sum_{n \in B_{\text{maj}}} q_n z_n\Big], \tag{4}$$

an attraction towards the slot and a softmax-weighted repulsion from the batch's majority, both scaled by $1/\tau$ and both saturating as the pair is already resolved ($p_j \to 1$). Gradient flows through z_a and through the majority views only.

Degenerate cases. The minority term needs two filled slots, so it is zero until two distinct minority samples have been drawn, the first step or two. It is zero for the entire run if the split holds a single minority example, $m = 1$, since $J(a)$ is then always empty; our smallest split has $m = 35$, so this does not arise here. And it is zero in any batch that draws no minority sample, at $\pi = 0.01$, $B = 32$ a case that recurs throughout training; what the bank guarantees is that whenever an anchor *does* appear, its positive set is the whole class.

Majority term (Eq. 3). A majority view is contrasted against its augmentation partner, with the batch's remaining majority views in the denominator (partner included, anchor excluded, as in NT-Xent) and every filled bank slot alongside them at weight w.

The weight w (Eq. 3). Entering all $\hat{m}$ slots at full weight would over-represent the minority in the denominator, because natural sampling would have supplied $\mathbb{E}|B_{\min}| = 2B\pi$ minority views, not $\hat{m}$. Taking the expectation of the contribution the batch's own minority views would have made, under sampling at rate π from the minority training set,

$$\mathbb{E}\Big[\sum_{n\in B_{\min}} e^{z_i\cdot z_n/\tau}\Big] = 2B\pi\;\mathbb{E}_j\big[e^{z_i\cdot \tilde{z}_j/\tau}\big] = \frac{2B\pi}{\hat{m}}\sum_{j=1}^{\hat{m}} e^{z_i\cdot \tilde{z}_j/\tau}, \tag{5}$$

which is the definition of w: an inclusion-probability correction that leaves the expected denominator equal to what natural sampling would have produced while removing its variance.

Aggregation. $\mathcal{L} = (\mathcal{L}_{\min} + \mathcal{L}_{\text{maj}})/2B$ divides both sums by the total number of anchors in the batch, a constant. Normalising per class instead, $\mathcal{L}_{\min}/|B_{\min}|$, ties each minority anchor's weight to a binomial draw: the fewer minority instances a batch happens to contain, the harder each of them pulls, so the minority gradient scale swings with the batch composition alone. Fixing the weight at $1/2B$ is what makes the minority gradient the same size whether the batch drew one minority sample or many.

Cost. Per batch, $O((2B)^2 d)$ for the majority similarity matrix and $O(2B\hat{m}d)$ for the bank products; the bank state is $m \times d$ floats, from 18 KB at PneumoniaMNIST 1% to 0.6 MB at its native rate.

A.2 OrthoProto

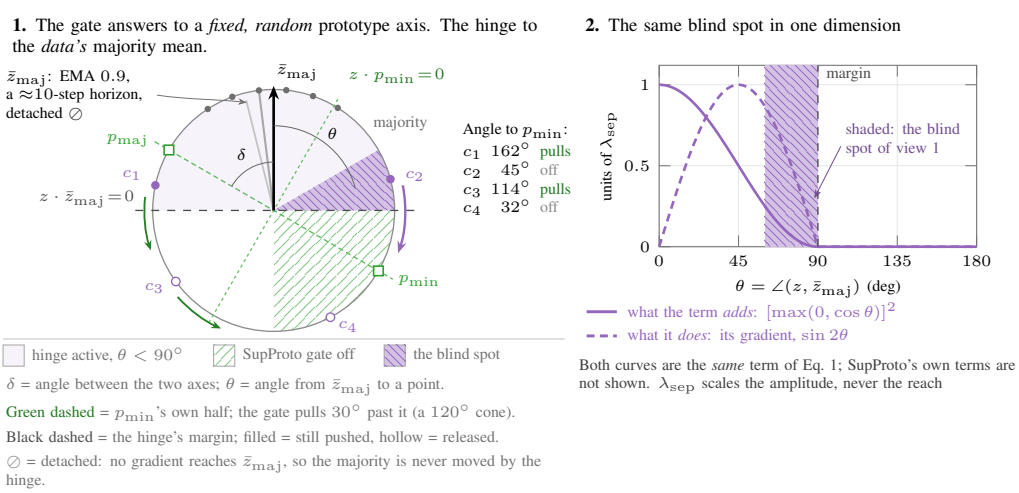

Figure 3: **OrthoProto. (1)** SupProto's gate answers to the prototype axis fixed before training while the hinge answers to the data's majority mean; the hatched region between them is SupProto's blind spot. The four marked embeddings span the cases: pull and hinge agree (c_1); the blind spot, where the gate has released a sample the hinge still pushes (c_2); the converse, where the hinge is silent but SupProto keeps pulling (c_3); and the intended end state, separated and left alone (c_4). Satisfying one criterion says nothing about the other because they answer to different targets. **(2)** The same blind spot in one dimension, with the penalty of Eq. 1 and the force $\lambda_{\text{sep}} \sin 2\theta$ it exerts.

OrthoProto is Eq. 1: the full SupProto objective plus one hinge term.

- $\bar{z}_{\mathrm{maj}} \in \mathbb{S}^{d-1}$: an exponential moving average of the majority's batch-mean direction, updated once per step as $\bar{z}_{\mathrm{maj}} \leftarrow \mathrm{norm}\big(0.9\,\bar{z}_{\mathrm{maj}} + 0.1\,\mathrm{mean}_{i \in B_{\mathrm{maj}}} z_i\big)$, detached from the graph. It is the loss's only state, one d-vector.
- $\max(0, z_a \cdot \bar{z}_{\mathrm{maj}})^2$: the squared hinge. It is non-zero exactly when the minority view a is within $90°$ of the majority-centre direction, grows quadratically in the cosine, and its force on the sphere is $\lambda_{\mathrm{sep}} \sin 2\theta$ where θ is the angle to $\bar{z}_{\mathrm{maj}}$ (Figure 3, panel 2): maximal at $45°$, zero at orthogonality. Past $90°$ the term is identically zero, so it never compresses a separated minority.
- $\frac{1}{|B_{\mathrm{min}}|}$: the hinge is averaged per minority anchor (unlike CompMin's $1/2B$ aggregation); with $\lambda_{\mathrm{sep}} = 1$ its magnitude is comparable to the prototype term it corrects. Batches with no minority view contribute no hinge term.
- λ_{sep}: the only new hyperparameter, 1 in every reported cell. $\lambda_{\mathrm{sep}} = 0$ is verified to reproduce SupProto exactly.

Gradient flows through the minority views only; $\bar{z}_{\mathrm{maj}}$ is a target, so majority embeddings feel the hinge only indirectly, through the encoder. Two measured properties justify the design. First, the direction matters: the hinge answers to where the majority actually is, and under SupProto alone the minority ends training inside the penalised hemisphere in every extreme cell (Table 9) even though the prototype gate is satisfied, which is exactly the blind spot of Figure 3. Second, the pairing matters: the hinge transplanted onto SupMin or CompMin at the same λ collapses the representation (the term supplies only repulsion from $\bar{z}_{\mathrm{maj}}$, and without the prototype's counterbalancing attraction the encoder satisfies it degenerately), so the hinge is reported as part of the SupProto+hinge pair, never as a standalone regulariser.

A.3 The separation gradient, derived in full

This section derives the tangential gradient of the separation hinge of Eq. 1: the Euclidean gradient of the squared angular hinge, its projection onto the tangent space of the unit sphere, and the closed form of its norm. The result is what licenses the claim in Section 4.1 that the hinge is bounded at both ends of its angular support.

Setup

Embeddings are unit vectors by construction (Section 3), so they live on

$$\mathbb{S}^{d-1} = \big\{ z \in \mathbb{R}^d : \|z\| = 1 \big\} \tag{6}$$

Write $c = \bar{z}_{\mathrm{maj}}$ for the running majority centre of Eq. 1, which is renormalised at every step and therefore also satisfies $\|c\| = 1$. Let $\theta \in [0, \pi]$ be the angle between z and c. Since both are unit vectors,

$$z^\top c = \|z\|\,\|c\| \cos\theta = \cos\theta \tag{7}$$

The separation penalty carried by a single minority anchor is

$$h(z) = \big[\max\big(0, z^\top c\big)\big]^2 = \begin{cases} (z^\top c)^2 = \cos^2\theta, & 0 \le \theta < \frac{\pi}{2}, \\ 0, & \frac{\pi}{2} \le \theta \le \pi. \end{cases} \tag{8}$$

Two observations for what follows. First, c is detached from the computational graph, so every derivative below treats it as a constant vector and no gradient reaches the majority centre. Second, the penalty is applied only to minority anchors, so z below always denotes a minority embedding.

Step 1: the Euclidean gradient

On the active branch $\theta < \pi/2$ the penalty is the square of a linear form. Applying the chain rule to $u \mapsto u^2$ composed with $z \mapsto z^\top c$, and using $\nabla_z (z^\top c) = c$,

$$\nabla_z h = 2\,\big(z^\top c\big)\,\nabla_z \big(z^\top c\big) = 2\,\big(z^\top c\big)\,c = 2\cos\theta\,c \tag{9}$$

On the inactive branch h is identically zero, so $\nabla_z h = 0$. The two branches agree at the boundary: as $\theta \to \pi/2^-$, Eq. 9 tends to $2\cos(\pi/2)c = 0$. The squared hinge is therefore continuously differentiable across its own margin, which an unsquared hinge is not.

STEP 2: WHY THE EUCLIDEAN GRADIENT IS THE WRONG OBJECT

The gradient above assumes z can move freely in any direction in space. However, z is constrained to the surface of a unit sphere. Because the embedding must stay on the sphere, it can only move tangentially: any valid direction of motion at point z must be perpendicular to z itself, since otherwise the movement would pull z off the sphere. This set of allowable directions is the tangent space:

$$T_z\mathbb{S}^{d-1} = \{v \in \mathbb{R}^d : v^\top z = 0\} \tag{10}$$

Any part of $\nabla_z h$ pointing along z itself is radial and must be removed.

STEP 3: THE TANGENTIAL PROJECTION

To remove the radial component of the gradient, we apply the projection matrix $P_z = I - zz^\top$. Because z is a unit vector, this matrix subtracts exactly the movement pointing along the radius. Applying it to the Euclidean gradient of Eq. 9 yields the surface gradient:

$$\nabla_{\mathbb{S}} h = P_z(2\cos\theta\, c) = (I - zz^\top)(2\cos\theta\, c) \tag{11}$$

To see how this multiplies out, distribute the terms. Multiplying by the identity leaves the vector unchanged, giving $I(2\cos\theta\, c) = 2\cos\theta\, c$. For the second term, because $z^\top c$ is a scalar, we can group it as $zz^\top c = z(z^\top c)$. The full expansion is therefore:

$$\begin{aligned}\nabla_{\mathbb{S}} h &= I(2\cos\theta\, c) - zz^\top(2\cos\theta\, c) \\ &= 2\cos\theta\, c - 2\cos\theta\, z(z^\top c) = 2\cos\theta\, c - 2\cos\theta(\cos\theta)z = 2\cos\theta(c - \cos\theta\, z)\end{aligned} \tag{12}$$

Note that $c - \cos\theta\, z$ is exactly the component of c that is orthogonal to z. The hinge therefore acts along the tangential direction that points directly at the majority centre. Because the optimiser takes a descent step $z \leftarrow z - \eta\nabla_{\mathbb{S}} h$, it pushes the anchor in the opposite direction, away from the majority centre, increasing the angle θ.

STEP 4: THE NORM

Because $2\cos\theta$ is a non-negative scalar,

$$\|\nabla_{\mathbb{S}} h\| = 2\cos\theta\, \|c - \cos\theta\, z\| \tag{13}$$

Expanding the remaining norm and using $\|c\| = \|z\| = 1$ together with Eq. 7,

$$\begin{aligned}\|c - \cos\theta\, z\|^2 &= c^\top c - 2\cos\theta\, c^\top z + \cos^2\theta\, z^\top z \\ &= 1 - 2\cos^2\theta + \cos^2\theta \\ &= 1 - \cos^2\theta = \sin^2\theta\end{aligned} \tag{14}$$

On $0 \le \theta < \pi/2$ we have $\sin\theta \ge 0$, so $\|c - \cos\theta\, z\| = \sin\theta$. Substituting into Eq. 13 and applying $2\sin\theta\cos\theta = \sin(2\theta)$,

$$\|\nabla_{\mathbb{S}} h\| = 2\cos\theta\sin\theta = \sin(2\theta), \qquad 0 \le \theta < \frac{\pi}{2}, \tag{15}$$

which is the force quoted in Figure 3. Outside that interval the penalty is inactive and the norm is zero.

STEP 5: A SIMPLE BREAKDOWN

Finally, how the force of Eq. 15 behaves:

- At $\theta = 0$ the force vanishes, so exact alignment with the majority centre is a stationary point that the hinge alone cannot escape, though the SupProto and NT-Xent terms and stochastic batch composition move an anchor off that measure-zero configuration in practice.
- At $\theta = \pi/4$ the force peaks at $\sin(\pi/2) = 1$, pushing hardest halfway to the margin, where the anchor sits unambiguously on the majority side but can still be closer to its prototype.
- As $\theta \to \pi/2^-$ the force anneals smoothly to zero, so the term stops pushing exactly where its stated goal is met rather than compressing a class that has already separated.
- The separation weight λ_{sep} multiplies Eq. 15. It scales the amplitude of this force but leaves its angular support unchanged, which is why it acts as a pure strength control ($\lambda_{\text{sep}} = 1$ everywhere in this paper).

B METRIC IMPLEMENTATIONS

All prevalence-projected test metrics are computed from one forward pass over the full official test split per selection state. No balanced test subset is ever constructed. Prevalence-dependent metrics at a stated rate π' use importance weights $w_i = 1$ for positives and $w_i = n_+(1-\pi')/(\pi' n_-)$ for negatives, so the weighted prevalence is exactly π' and every test image is retained. AUPRC_{50} is average precision under these weights at $\pi' = 0.5$ (the field's conventional rate) and AUPRC_π at the training (deployment) rate. AUROC and ER-AUROC are computed on the raw scores and are unaffected by the weighting; we verified by recomputation from the archived predictions that the logged values equal the plain AUROC of the unweighted test split to four decimals. ER-AUROC is the standardised partial AUROC with $\text{FPR}_{\max} = 0.1$ (McClish, 1989), for which chance is 0.5. F1_π applies the probe's default (0.5) threshold under the deployment-rate weighting. Model selection everywhere is best validation AUROC on the balanced validation split, and the test set is never touched during training.

Figure 4 quantifies the reliability gap between the two retrieval lenses across the whole study: every dataset, every prevalence rung and every batch size, one point per (method, cell) with three or more matched seeds, 171 pairs in total. Each point is the ratio of the across-seed standard deviation of AUPRC_π to that of ER-AUROC, on identical runs, plotted against the cell's minority share. AUPRC_π is the noisier lens in 80% of all pairs, and the imbalance the paper studies makes it worse: the median ratio grows from near parity at native rates to $3.4\times$ at a 1% share. At three seeds, differences of that size cannot be resolved on AUPRC, which is why the tables lean on ER-AUROC for retrieval comparisons and read AUPRC only within a cell.

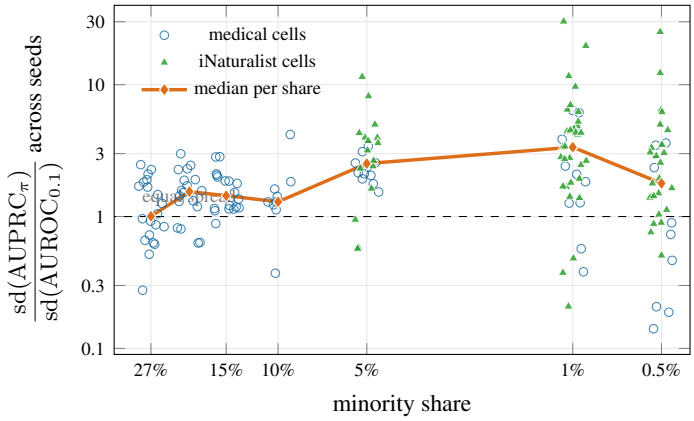

Figure 4: Reliability of AUPRC at deployment prevalence relative to ER-AUROC. Each point is one (method, cell) pair with matched seeds; its height is the ratio of the two metrics' across-seed standard deviations on identical runs, above the dashed line meaning AUPRC is the noisier lens. The solid line traces the median per prevalence. Both axes are logarithmic, and the minority share decreases to the right.

C DECISION-ANALYTIC DETAILS

Definition. Write $u \otimes v = uv/(uv + (1-u)(1-v))$ for odds-scale combination. Under label shift a score calibrated at the evaluation rate π_{eval} transports to a deployment prevalence π' by odds multiplication (Saerens et al., 2002), and Flores et al. (2025) show the cost-optimal rule then treats when the raw score exceeds $t(\pi') = c \otimes (\pi_{\text{eval}} \otimes (1-\pi'))$, earning there the prior-adjusted, cost-weighted accuracy

$$V(\pi') \;=\; \text{TPR}\big(t(\pi')\big) \cdot \big[(1-c) \otimes \pi'\big] \;+\; \text{TNR}\big(t(\pi')\big) \cdot \big[c \otimes (1-\pi')\big] \;\in\; [0,1], \qquad (16)$$

normalised so that a perfect classifier scores 1 at every prevalence in the band. The reported score is the average of Eq. 16 over the band, taken uniformly in log-odds,

$$S_{\mathrm{DCA}} \;=\; \mathbb{E}_{\mathrm{logit}(\pi')\sim\mathcal{U}[\mathrm{logit}\,a,\,\mathrm{logit}\,b]}\big[V(\pi')\big], \tag{17}$$

for which Flores et al. (2025) derive a closed form: up to a factor fixed by c and the band, S_{DCA} equals a log score evaluated on probabilities clipped to the range the band can reach. That equivalence is the only property of S_{DCA} we rely on.

Score computation. The band $[a, b]$ is discretised into 150 points uniform in log-odds. At each π' the threshold of Eq. 16 is applied to the sorted score array and TPR/TNR are read off exactly. The medical headline scores use each dataset's declared operating point (band and k below).

Calibration decomposition. $S^{\mathrm{cal}}_{\mathrm{DCA}}$ answers one question: *how much of a model's decision loss would disappear if its probabilities were perfectly recalibrated while its ranking stayed fixed?* We recalibrate with isotonic regression and score again. The gap $S^{\mathrm{cal}}_{\mathrm{DCA}} - S_{\mathrm{DCA}}$ is then the price of miscalibration, and whatever deficit remains after recalibration is a ranking problem that no calibration can fix.

Label orientation. On PneumoniaMNIST the statistical minority (our class 1) is the normal X-ray, while the acted-on class is pneumonia. Scores and labels are flipped before decision scoring, so the band is a prevalence range of the acted-on class, and the identity $(\text{positive} = A, c, [a, b]) \equiv (\text{positive} = B, 1 - c, [1 - b, 1 - a])$ makes the two orientations consistent.

Dataset	Band (acted-on class)	Cost k
BreastMNIST	1–25%	50 (biopsy threshold, D'Orsi et al., 2013; Burnside et al., 2012)
PneumoniaMNIST	5–20%	9 (test threshold, Ebell et al., 2018)
iNaturalist	not scored	no declarable operating point

Table 5: Operating points behind every reported S_{DCA} number. The medical headlines score at a single declared band and cost each; iNaturalist declares no operating point, so S_{DCA} is not reported there (Section 7.2). Breast band and cost follow the BI-RADS 3/4a biopsy boundary: a lesion with malignancy probability above roughly 2% warrants biopsy (D'Orsi et al., 2013; Fazeli et al., 2025; Burnside et al., 2012), and k=50 reproduces that boundary as the break-even probability $c = 1/(1 + k) \approx 2\%$; the pneumonia band brackets the radiographic-pneumonia rate in the source population (Neuman et al., 2011; Kermany et al., 2018) and its cost matches the empirically measured test threshold for community-acquired pneumonia, 9.5% [8.7, 10.5] (Ebell et al., 2018), whose break-even probability $c = 1/(1 + k) = 10\%$ the ratio k=9 reproduces. Both operating points are illustrations of the decision lens, not clinical recommendations: a deployment must re-derive its own band and cost, and none of these numbers should be used as a clinical reference.

D Training details

Contrastive stage. We train a ResNet-50 encoder from scratch. The projection head is a two-layer MLP producing 128-dimensional L2-normalized embeddings, and we fix the temperature at $\tau = 0.07$ across all contrastive losses. We use a standard batch size of 256. For the batch robustness study, this is reduced to 32 with the learning rate scaled linearly. Models are optimized using SGD with a momentum of 0.9, a peak learning rate of 0.0625 following a cosine decay schedule with 10 warm-up epochs, and a weight decay of 10^{-4}. We apply standard SimCLR augmentations at a 224×224 pixel resolution. Pretraining runs for 350 epochs on iNaturalist and 100 epochs on MedMNIST to exactly match the protocol of Mildenberger et al. (2025). We apply no class-balanced sampling during the contrastive stage; batch composition remains strictly binomial at the native training prevalence. Because the two dataset families employ different probe protocols (detailed below), absolute performance scores cannot be directly compared across the iNaturalist and MedMNIST boundary.

Probe stage. During evaluation, the encoder and its batch-norm statistics are strictly frozen. A linear probe is trained for 50 epochs using standard cross-entropy loss on a class-balanced pool. On iNaturalist we train two probes per run, identical in every setting (Adam, learning rate 3×10^{-4}, cosine schedule, batch size 1024) except the pool they train on, in order to analyze how the probe's label budget changes the comparison (Tables 3 and 4). The full-pool probe trains on the full balanced

training set and scores the frozen features with abundant minority labels; it is also the configuration under which the published balanced accuracies of Mildenberger et al. (2025) are reproducible, and restricting the pool drops performance by several points on identical frozen features. The thin probe trains on the balanced subset of the thinned split alone, so the pipeline consumes no label beyond the pretraining budget. For the MedMNIST datasets, the probe is trained exclusively on the balanced subset of the thinned pretraining split using SGD at standard dataset settings with a batch size of 1024.

Method hyperparameters and seeds. For CompMin, we set the bank momentum to $\mu = 0.5$ using per-example slot assignments and apply inclusion-probability bank weighting in the majority denominator as defined in Eqs. 2–3. For OrthoProto, we use an EMA rate of 0.9, a margin of 90°, and set the separation weight $\lambda_{\text{sep}} = 1$ consistently across all experiments. All runs report the average of seeds 2000, 2001, and 2002. To account for its higher variance, BreastMNIST additionally includes seeds 2003 and 2004 across every prevalence rung.

E Dataset construction

MedMNIST (Yang et al., 2023). We use the official train, validation, and test splits at a 224×224 pixel resolution. Class 1 serves as the statistical minority, representing malignant cases in BreastMNIST (native share 27%) and normal cases in PneumoniaMNIST (native share 26%). The thinned evaluation rungs subsample the minority of the training split exclusively, leaving the majority fixed. This training rate directly defines the deployment rate used for both the deployment-prevalence metrics and the decision-band centring. The validation set utilizes the official split balanced to the smaller class, ensuring that AUROC-based model selection remains entirely prevalence-free. An important structural asymmetry exists in this setup: the validation split is never thinned. At the lowest prevalence rungs, it consequently holds more labelled positive examples than the training set itself (for example, the 1% PneumoniaMNIST rung contains 35 training positives against 135 validation positives). The extreme imbalance regime therefore constrains representation learning rather than model selection or thresholding. While this matches the protocol established by Mildenberger et al. (2025), any deployment-focused interpretation of the thinned rungs must account for this characteristic.

iNaturalist 2021 (Van Horn et al., 2018). Three binary tasks by taxonomic prefix: Quercus vs Saxifragales (plants), Apidae vs Vespidae (insects), Artiodactyla vs Carnivora (mammals). The train directory is split 95/5 into train and balanced validation, and the official val split provides the test images. Cells fix the total training size and set the positive share to 5%, 1% or 0.5%, extending the construction of Mildenberger et al. (2025) one prevalence rung lower.

Table 6 collects the per-experiment statistics. Every task is binary (positive = minority = class 1). Thinning changes only the training composition; the full test split is shared by all rungs of a task, is identical across seeds and methods, and is never subsampled at evaluation time.

Task	Rung	Train N	Train pos.	Test N	Test pos.
BreastMNIST	native (27%)	546	147	156	42
	20%	498	99	156	42
	15%	469	70	156	42
PneumoniaMNIST	native (26%)	4708	1214	624	234
	20%	4367	873	624	234
	15%	4110	616	624	234
	10%	3882	388	624	234
	5%	3677	183	624	234
	1%	3529	35	624	234
	0.5%	3511	17	624	234
Plants (Quercus vs Saxifragales)	95:5	11196	559	1250	430
	99:1	11193	111	1250	430
	99.5:0.5	11196	55	1250	430
Insects (Apidae vs Vespidae)	95:5	9440	472	800	420
	99:1	9439	94	800	420
	99.5:0.5	9439	47	800	420
Mammals (Artiodactyla vs Carnivora)	95:5	14609	730	1090	550
	99:1	14609	146	1090	550
	99.5:0.5	14609	73	1090	550

Table 6: Per-experiment data statistics, two classes throughout, read from the archived run logs.

F Representation diagnostics

We log the five diagnostics, two of them introduced by Mildenberger et al. (2025), for every contrastive run. SAD measures the mean distance between a sample's two augmented views (Wang & Isola, 2020). SAA records the percentage of views whose nearest neighbour is their own second view. CAD calculates the mean within-class pairwise distance (Li et al., 2022). CAC tracks the percentage of a view's nearest 5% of neighbours that share its class. GPU evaluates the mean pairwise Gaussian potential. CAD yields low values for both tightly clustered and fully collapsed classes. SAA and CAC accompany it specifically to disambiguate these two states.

All five metrics are computed using the authors' released implementation. This codebase reports each diagnostic per class and returns the raw Gaussian potential rather than its logarithm. A GPU score of 1.000 therefore indicates a class entirely collapsed to a single point while lower values reflect a more uniform distribution. The official implementation accepts any pairwise distance matrix as an argument. We supply the cosine distance computed on the L2-normalised embeddings. Because SAA and CAC depend strictly on neighbour rankings and GPU does not rely on this matrix at all, these three metrics remain completely unaffected by this choice. SAD and CAD simply operate on a different scale compared to the ℓ_2 values plotted in the original paper.

We augment this suite with the geometric violation diagnostic discussed in the main text. This metric calculates the fraction of minority training embeddings located within 90° of the EMA-estimated majority-centre direction at the end of the contrastive stage. We also record their mean cosine similarity to this centre. All geometric properties are measured exclusively on the projection space at the end of the contrastive stage. The linear probe instead evaluates the frozen backbone. These tables therefore constrain the projection space alone. This architectural distinction governs the downstream reading: the probe scores the backbone, not the projection.

Figures 5 and 6 report all five diagnostics, one panel per metric: cells on the vertical axis, the metric on the horizontal axis, one marker per method, open for the majority class and filled for the minority. The medical panel carries all seven contrastive methods and the iNaturalist panel the four minority-aware methods of the main board.

Reading the medical panels (Figure 5). The GPU panel carries the collapse story. Mildenberger et al. (2025) proved that contrastive collapse is majority-driven and whole-space, and the panel shows exactly that mode: on BreastMNIST every prototype-anchored objective (SupProto, Ortho-Proto, SupCon, KCL, PaCo) sits at 1.0 for both classes, a class collapsed to a point, while SupMin

and CompMin, whose majority term is instance-level by design, never do; on PneumoniaMNIST the in-batch SupCon, KCL and PaCo carry a collapsed majority at every rung against 0.04 to 0.2 for the SupMin family, and CompMin keeps the most dispersed minority in every BreastMNIST cell and at pneumonia 1%. The SAD, CAD and SAA panels then illustrate why Mildenberger et al. (2025) distrust the classical metrics: wherever GPU reads 1.0, SAD and CAD read near zero and SAA collapses too, because a space compressed to a point makes every distance small and no view can keep its own partner as nearest neighbour, so a tightly clustered class and a destroyed one are indistinguishable to these scores. CAC is the exception: it is the one panel whose ordering resembles the accuracy boards, which is exactly the predictive-validity result of Table 7. And none of it dictates downstream quality on its own: at BreastMNIST 20% OrthoProto ends pretraining completely collapsed ($\mathrm{GPU} = 1.000$) yet is the best method in the cell, because the probe reads the backbone, not the projection.

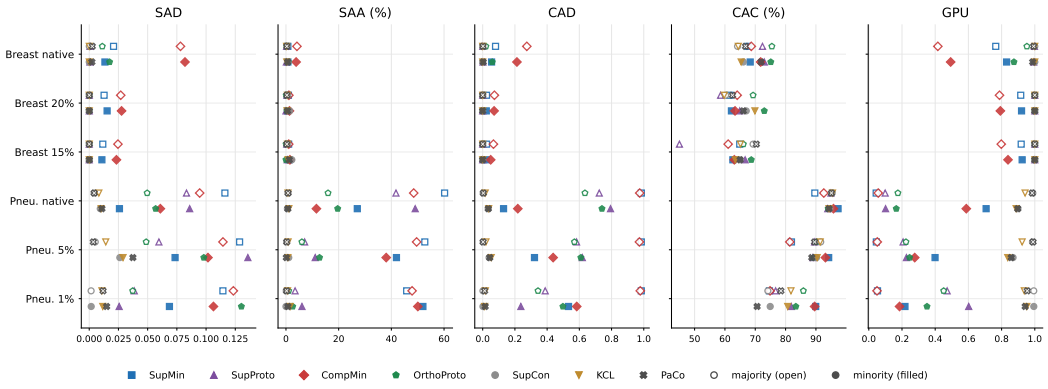

Figure 5: The five representation diagnostics on the medical cells (mean over the matched seeds: five at BreastMNIST, three at PneumoniaMNIST; batch 256). One marker per method; open = majority class, filled = minority class. SAD and CAD are cosine distances, GPU is the Gaussian potential (1.0 = a class collapsed to a point), SAA and CAC are percentages.

Reading the iNaturalist panels (Figure 6). The medical failure mode does not transfer: GPU stays below 0.25 and SAA above 65% in every cell, so nothing collapses on iNaturalist. The geometric failure there is the prototype conflict of Section 7.2, which lives in *where* the minority sits (the violation rate of Table 8), not in how tightly anything clusters. Two patterns are still worth reading off. SupMin keeps the lowest SAD everywhere, its views embed closest, without that translating into a uniform accuracy lead, another reminder that the classical metrics rank geometry rather than outcomes. And CAC falls as the minority share falls in step with the accuracy boards, the same behaviour that makes it the one defensible global proxy in Table 7.

Table 7 analyzes the predictive validity of the representation diagnostics across all contrastive runs. The table reports the linear R^2 and Spearman ρ correlations between each end-of-pretraining metric and the final test AUROC. Following Mildenberger et al. (2025) CAC emerges as the only diagnostic whose global correlation survives on this benchmark. The canonical SAD, CAD, and GPU metrics prove weak as global predictors, and the SAA metric performs similarly poorly here. Our geometric violation rate is likewise a weak global predictor because its role remains strictly mechanistic rather than general. Furthermore all within-cell correlations remain modest. This confirms that no single geometric metric can reliably rank methods inside a fixed task.

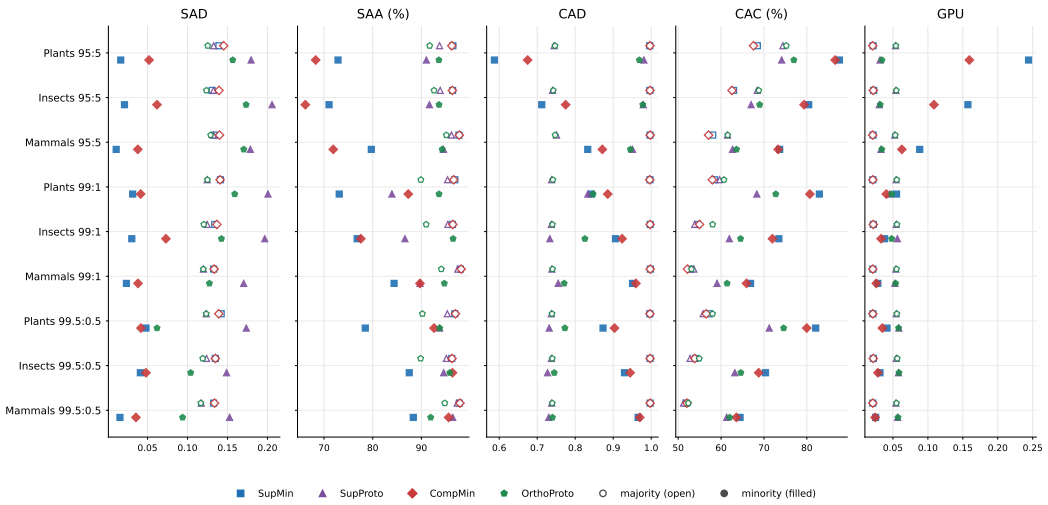

Figure 6: The five representation diagnostics on the iNaturalist cells (mean over seeds 2000–2002, batch 256), same layout as Figure 5.

	global R^2 vs AUROC				
Metric	pooled	medical	iNat	pooled ρ	mean within-cell $\lvert\rho\rvert$
SAD (majority)	0.242	0.31	0.105	0.401	0.367
SAD (minority)	0.131	0.303	0.001	0.429	0.359
SAA (majority)	0.138	0.284	0.042	0.317	0.365
SAA (minority)	0.093	0.198	0.1	0.29	0.291
CAD (majority)	0.311	0.386	0.0	0.446	0.326
CAD (minority)	0.173	0.295	0.05	0.391	0.331
CAC (majority)	0.342	0.772	0.635	0.629	0.441
CAC (minority)	0.613	0.864	0.681	0.852	0.36
GPU (majority)	0.325	0.391	0.001	-0.414	0.312
GPU (minority)	0.243	0.38	0.165	-0.346	0.323
Violation rate (ours)	0.246	0.328	0.049	-0.489	0.42

Table 7: Predictive validity of representation diagnostics. The columns report linear R^2 and Spearman ρ correlations of each metric against test AUROC. The final column displays the mean absolute within-cell correlation computed over every cell.

G PAIRED DIFFERENCES AND ADDITIONAL RESULT TABLES

This section provides the per-cell conflict percentages and paired-difference intervals supporting Sections 7.2 and 7.2.2. It also includes the paired thin-probe differences behind the matched-budget reading of Table 4.

Measuring the violation rate. We compute the violation rate once per run at the end of the contrastive stage using the training split. Every training image is embedded by the frozen encoder and projection head before being normalised to the unit sphere without any training augmentations. The majority mean direction $\mu_{\mathrm{maj}} = \bar{z}_{\mathrm{maj}}/\|\bar{z}_{\mathrm{maj}}\|$ is the normalised average of the majority-class embeddings. A minority embedding z_i violates when $z_i^\top \mu_{\mathrm{maj}} > 0$. This condition means the embedding lies within 90° of the majority centre and therefore falls inside the majority's half-space. The violation rate represents the overall fraction of minority training embeddings that meet this geometric condition. A rate of 0% indicates the two classes occupy entirely opposite hemispheres. A rate of 100% indicates the entire minority ends pretraining trapped in majority territory and leaves a linear probe little room for separation. The 90° threshold matches the exact margin enforced by the OrthoProto hinge in Eq. 1. Table 8 therefore reports the precise quantity that this objective penalises. This rate serves strictly as a mechanistic diagnostic rather than a model-selection metric. It does not correlate globally with test AUROC across runs (Table 7); only CAC does (R^2 0.68 iNaturalist, 0.86 medical, over all 537 contrastive runs), reproducing the headline claim of Mildenberger et al. (2025) under the corrected probe.

Task	Cell	SupMin	SupProto	CompMin	OrthoProto
Plants	95:5	0.0	5.9	0.0	3.0
Plants	99:1	0.3	47.2	0.0	5.7
Plants	99.5:0.5	52.1	97.0	0.0	4.3
Insects	95:5	34.0	7.3	0.1	3.1
Insects	99:1	0.0	100.0	0.4	3.9
Insects	99.5:0.5	31.2	100.0	0.0	2.8
Mammals	95:5	30.8	12.8	0.1	5.4
Mammals	99:1	0.2	79.5	0.0	3.0
Mammals	99.5:0.5	0.0	99.5	0.5	2.3

Table 8: Quantifies the geometric conflict by reporting the percentage of minority training embeddings that end pretraining within 90° of the majority-centre direction at batch 256 averaged across seeds 2000 to 2002. Under SupProto this fraction climbs toward 100% as the minority share falls. OrthoProto's hinge strictly maintains this rate between 3% and 6% across all cells. Paired per-cell damage and repair intervals are available in Table 9.

Task	Cell	SupProto − SupMin	OrthoProto − SupProto	SupProto viol. rate
Plants	95:5	-0.26 [-1.46, +0.93]	+0.86 [-0.26, +1.99]	6%
Plants	99:1	-1.32 [-1.62, -1.03]*	+0.11 [-0.97, +1.18]	47%
Plants	99.5:0.5	-1.88 [-2.56, -1.20]*	+1.08 [+0.42, +1.74]*	97%
Insects	95:5	-0.44 [-1.44, +0.57]	+1.43 [-0.15, +3.01]	7%
Insects	99:1	-5.59 [-6.80, -4.37]*	+7.58 [+5.78, +9.37]*	100%
Insects	99.5:0.5	-3.18 [-5.06, -1.30]*	+5.49 [+4.38, +6.59]*	100%
Mammals	95:5	-0.19 [-1.69, +1.31]	+1.02 [+0.08, +1.96]*	13%
Mammals	99:1	-1.71 [-4.07, +0.65]	+1.15 [+0.07, +2.22]*	80%
Mammals	99.5:0.5	-1.14 [-2.01, -0.27]*	+1.00 [-2.42, +4.42]	100%

Table 9: Quantifies the damage caused by SupProto alongside OrthoProto's repair and the geometric conflict for each iNaturalist cell at batch 256. The left two columns report paired per-seed AUROC differences scaled by 100 alongside 95% t-intervals. An asterisk marks any interval excluding zero. The right column displays the fraction of minority training embeddings that end pretraining within 90° of the majority-centre direction under SupProto averaged across three seeds. OrthoProto consistently restricts this same quantity strictly between 3% and 6% in every evaluated cell.

		p_B (min. views/batch)		Δ(bs32 − bs256)		CompMin − SupMin at bs32	
Task	Cell	bs256	bs32	SupMin	CompMin	AUROC	ER-AUROC
Plants	99:1	5.12	0.64	-1.6	+0.7	+1.5 [+0.5, +2.6]*	+5.8 [+4.0, +7.7]*
Insects	99:1	5.12	0.64	-3.8	+0.1	+3.7 [+0.0, +7.3]*	+6.0 [-1.3, +13.4]
Mammals	99:1	5.12	0.64	-1.7	-0.6	+0.2 [-0.8, +1.2]	+1.0 [+0.4, +1.6]*
Plants	99.5:0.5	2.56	0.32	-2.4	-0.6	+2.0 [+0.8, +3.2]*	+5.2 [+2.1, +8.4]*
Insects	99.5:0.5	2.56	0.32	-5.1	+0.4	+5.0 [+2.8, +7.3]*	+6.9 [+4.2, +9.6]*
Mammals	99.5:0.5	2.56	0.32	-4.2	-1.5	+2.3 [+0.8, +3.7]*	+3.9 [+0.6, +7.3]*

Table 10: Details the performance impact of shrinking the contrastive batch from 256 to 32. The variable $p_B = 2B\pi$ denotes the expected number of minority views per batch. The Δ columns report the mean paired per-seed AUROC change between batch sizes 256 and 32 for each evaluated loss. The rightmost columns present the paired CompMin minus SupMin difference at batch 32 alongside 95% t-intervals evaluated at $n = 3$. An asterisk indicates an interval that strictly excludes zero. All reported values are scaled by 100.

Task	Cell	OrthoProto − SupProto	OrthoProto − SupMin	CompMin − SupMin
Plants	95:5	+0.4 [-1.1, +1.9]	+0.6 [-1.1, +2.2]	-0.0 [-1.3, +1.2]
Insects	95:5	+1.2 [-1.7, +4.0]	+0.8 [-0.1, +1.6]	-0.8 [-1.1, -0.5]*
Mammals	95:5	+1.3 [-0.9, +3.5]	+1.6 [-1.8, +5.0]	+0.2 [-4.1, +4.5]
Plants	99:1	-2.4 [-8.5, +3.7]	-2.4 [-3.5, -1.4]*	+1.5 [-2.9, +5.9]
Insects	99:1	+11.0 [+7.7, +14.3]*	+2.5 [+0.8, +4.2]*	+0.6 [-1.5, +2.7]
Mammals	99:1	-4.8 [-8.6, -1.0]*	-2.5 [-4.1, -1.0]*	+1.4 [-1.1, +4.0]
Plants	99.5:0.5	+0.8 [-5.3, +6.9]	-4.4 [-7.9, -0.8]*	+1.1 [-2.1, +4.3]
Insects	99.5:0.5	+8.8 [+5.7, +11.9]*	+1.7 [-3.5, +6.9]	-0.4 [-4.6, +3.7]
Mammals	99.5:0.5	-4.1 [-10.0, +1.9]	-4.9 [-5.9, -4.0]*	+1.1 [-5.8, +8.1]

Table 11: Paired per-seed AUROC differences under the thin probe (Table 4), with 95% t-intervals (n=3); * excludes zero. All values $\times 100$.

H FULL RESULT TABLES

Per-cell tables with all reported metrics for every cell of the study. The medical decision-score columns use each dataset's declared operating point; the operating points are in Table 5.

Method	AUROC	ER-AUROC	BAcc	AUPRC$_\pi$	F1$_\pi$
SupMin	96.3±0.3	87.3±0.9	89.9±0.5	68.8±0.8	47.7±2.5
SupProto	96.0±0.5	87.9±1.1	89.5±0.8	71.4±4.3	45.6±2.7
CompMin	96.4±0.2	88.5±0.6	90.6±0.5	71.3±0.3	48.0±1.2
OrthoProto	**96.9±0.1**	**89.7±0.2**	**91.3±0.2**	**75.6±0.7**	**49.6±5.2**

Table 12: Plants 95:5, batch 256. All metrics ×100, mean±sd over matched seeds. Bold: best mean.

Method	AUROC	ER-AUROC	BAcc	AUPRC$_\pi$	F1$_\pi$
SupMin	**94.0±0.3**	**81.8±0.8**	**86.8±0.3**	**30.3±3.4**	10.2±0.9
SupProto	92.7±0.3	79.0±0.4	84.8±0.5	28.4±4.9	9.9±0.5
CompMin	93.3±0.1	80.7±0.5	85.5±0.1	29.0±3.6	10.1±2.1
OrthoProto	92.8±0.7	80.0±2.0	85.5±1.1	30.3±10.4	**10.5±1.7**

Table 13: Plants 99:1, batch 256. All metrics ×100, mean±sd over matched seeds. Bold: best mean.

Method	AUROC	ER-AUROC	BAcc	AUPRC$_\pi$	F1$_\pi$
SupMin	93.0±0.5	**79.6±0.7**	85.4±0.7	14.2±0.6	5.3±0.2
SupProto	91.1±0.3	73.2±0.9	83.9±0.4	10.5±2.7	4.6±0.6
CompMin	**93.2±0.2**	79.5±1.4	**85.6±0.3**	17.5±1.2	**5.4±0.6**
OrthoProto	92.2±0.5	78.6±1.1	84.6±0.7	**18.5±3.1**	5.2±0.5

Table 14: Plants 99.5:0.5, batch 256. All metrics ×100, mean±sd over matched seeds. Bold: best mean.

Method	AUROC	ER-AUROC	BAcc	AUPRC$_\pi$	F1$_\pi$
SupMin	92.6±0.3	82.7±0.9	84.1±1.1	66.5±2.2	32.0±4.7
SupProto	92.2±0.3	82.8±1.7	83.2±0.3	67.9±4.0	35.6±2.4
CompMin	92.3±0.5	82.5±0.9	84.2±0.6	67.0±0.5	40.7±7.9
OrthoProto	**93.6±0.3**	**83.5±1.2**	**85.6±0.7**	**69.4±2.9**	**41.3±1.8**

Table 15: Insects 95:5, batch 256. All metrics ×100, mean±sd over matched seeds. Bold: best mean.

Method	AUROC	ER-AUROC	BAcc	AUPRC$_\pi$	F1$_\pi$
SupMin	86.6±0.7	72.2±0.7	78.4±1.4	26.0±1.2	6.8±0.6
SupProto	81.0±0.9	64.5±0.2	73.1±0.5	10.0±0.7	5.5±0.4
CompMin	86.3±0.6	72.0±0.9	77.9±1.2	26.7±5.8	7.1±0.7
OrthoProto	**88.6±1.1**	**75.7±0.6**	**79.8±1.9**	**35.7±1.8**	**7.7±1.2**

Table 16: Insects 99:1, batch 256. All metrics ×100, mean±sd over matched seeds. Bold: best mean.

Method	AUROC	ER-AUROC	BAcc	AUPRC$_\pi$	F1$_\pi$
SupMin	84.0±0.5	68.4±1.0	75.4±0.6	18.5±6.4	3.0±0.3
SupProto	80.8±0.3	63.6±0.9	72.8±0.8	5.9±1.3	2.8±0.6
CompMin	83.5±0.2	67.8±0.9	75.4±0.2	15.0±5.8	3.2±0.1
OrthoProto	**86.3±0.2**	**70.8±1.0**	**78.1±1.0**	**19.9±2.7**	**3.3±0.3**

Table 17: Insects 99.5:0.5, batch 256. All metrics $\times 100$, mean±sd over matched seeds. Bold: best mean.

Method	AUROC	ER-AUROC	BAcc	AUPRC$_\pi$	F1$_\pi$
SupMin	86.9±1.0	74.5±1.2	78.2±1.0	50.8±4.9	28.9±1.8
SupProto	86.7±0.8	73.7±0.3	77.9±1.4	45.6±2.6	**30.1±4.7**
CompMin	86.7±0.3	74.0±0.4	78.3±0.6	48.3±4.6	26.4±0.4
OrthoProto	**87.7±0.4**	**75.8±0.9**	**79.1±0.8**	**51.5±2.9**	29.5±1.6

Table 18: Mammals 95:5, batch 256. All metrics $\times 100$, mean±sd over matched seeds. Bold: best mean.

Method	AUROC	ER-AUROC	BAcc	AUPRC$_\pi$	F1$_\pi$
SupMin	**83.4±0.1**	69.1±0.2	**75.6±0.5**	17.3±4.9	**6.9±1.3**
SupProto	81.7±0.9	67.4±0.8	74.0±1.5	18.7±3.2	5.9±0.6
CompMin	82.5±0.1	68.4±0.2	74.2±1.2	16.1±4.8	6.5±1.4
OrthoProto	82.9±0.8	**69.3±1.2**	74.4±0.8	**19.4±3.3**	6.5±0.8

Table 19: Mammals 99:1, batch 256. All metrics $\times 100$, mean±sd over matched seeds. Bold: best mean.

Method	AUROC	ER-AUROC	BAcc	AUPRC$_\pi$	F1$_\pi$
SupMin	**81.5±0.6**	67.2±1.3	**73.6±0.8**	**12.1±2.3**	2.9±0.7
SupProto	80.3±0.7	65.7±0.6	72.0±1.1	12.1±0.9	2.9±0.3
CompMin	81.0±0.7	**67.2±1.6**	73.0±0.5	11.7±2.4	**3.1±0.6**
OrthoProto	81.3±2.0	66.9±2.1	73.4±1.5	9.9±2.9	2.9±0.4

Table 20: Mammals 99.5:0.5, batch 256. All metrics $\times 100$, mean±sd over matched seeds. Bold: best mean.

Method	AUROC	ER-AUROC	BAcc	AUPRC$_\pi$	F1$_\pi$	S_{DCA}	$S_{\mathrm{DCA}}^{\mathrm{cal}}$
Logit adj.	72.9±3.6	56.9±4.8	62.6±7.0	50.9±8.6	52.6±5.7	71.3±8.9	80.1±1.9
BCL	71.4±5.6	55.5±5.2	67.3±5.6	47.2±9.9	50.9±5.0	72.6±2.7	79.8±1.8
SupCon	62.5±7.8	54.3±2.6	56.8±6.9	40.5±5.8	41.8±11.4	70.8±0.3	77.5±1.8
KCL	62.6±7.9	53.9±3.4	57.7±5.5	40.8±8.4	36.8±14.5	70.7±0.2	77.5±1.9
PaCo	70.3±7.5	57.3±4.7	63.8±7.5	49.7±9.8	50.5±7.9	72.3±1.6	79.6±2.6
SupMin	70.8±2.1	57.9±3.9	61.3±5.9	49.7±6.5	50.6±2.9	74.8±1.5	79.1±0.6
SupProto	79.3±2.5	66.0±3.6	70.6±4.5	64.2±3.4	59.8±4.8	74.0±2.4	82.3±1.2
CompMin	76.8±1.5	63.5±2.0	69.6±2.0	59.6±3.4	55.7±1.8	75.7±0.9	81.4±0.8
OrthoProto	**83.0±3.0**	**70.6±3.8**	**74.2±2.4**	**70.0±4.7**	**62.5±1.8**	**77.5±3.1**	**84.0±1.6**

Table 21: BreastMNIST native, batch 256. All metrics $\times 100$, mean±sd over matched seeds. Bold: best mean.

Method	AUROC	ER-AUROC	BAcc	AUPRC$_\pi$	F1$_\pi$	S_{DCA}	$S_{\mathrm{DCA}}^{\mathrm{cal}}$
Logit adj.	67.1±5.0	54.3±1.6	59.8±5.3	35.3±4.8	39.8±5.7	69.9±3.5	78.6±2.1
BCL	63.6±6.8	52.9±2.8	54.0±3.8	30.4±5.4	37.1±4.7	72.8±2.1	77.6±2.5
SupCon	61.8±6.0	52.8±2.7	57.1±3.5	31.1±6.1	35.7±4.4	70.5±0.1	77.4±1.8
KCL	59.2±3.4	52.1±3.1	56.0±4.0	28.5±5.0	31.3±4.0	70.5±0.1	76.3±1.1
PaCo	61.5±4.8	53.5±1.5	57.5±3.6	30.6±2.6	32.8±4.6	70.7±0.2	77.5±1.3
SupMin	59.1±2.9	51.4±0.6	56.4±2.8	26.1±1.4	34.4±2.0	70.7±0.4	75.8±0.7
SupProto	67.3±6.4	56.5±2.9	59.8±4.4	38.3±7.0	40.2±5.9	70.8±0.3	78.4±1.9
CompMin	64.5±5.4	54.4±3.5	58.8±5.3	33.6±6.6	38.2±2.8	**73.2±1.8**	77.2±1.2
OrthoProto	**77.0±3.3**	**61.9±3.9**	**69.8±3.1**	**49.4±5.5**	**50.3±6.4**	72.7±1.5	**81.6±1.3**

Table 22: BreastMNIST 20%, batch 256. All metrics ×100, mean±sd over matched seeds. Bold: best mean.

Method	AUROC	ER-AUROC	BAcc	AUPRC$_\pi$	F1$_\pi$	S_{DCA}	$S_{\mathrm{DCA}}^{\mathrm{cal}}$
Logit adj.	62.8±4.5	52.1±2.3	54.4±4.8	22.7±2.0	30.5±3.1	**72.0±1.7**	77.7±1.7
BCL	64.0±3.8	53.4±1.4	57.1±5.4	24.9±2.6	30.5±4.5	70.4±4.6	77.5±1.6
SupCon	65.3±3.6	54.2±1.4	62.2±1.9	27.5±3.9	27.5±5.7	70.6±0.1	78.3±1.1
KCL	61.0±5.9	54.1±2.9	55.7±5.8	24.7±6.0	27.6±2.4	70.7±0.1	77.4±1.4
PaCo	64.5±5.1	53.4±3.0	57.9±3.6	25.9±6.2	27.0±4.6	70.7±0.1	77.7±1.3
SupMin	59.7±3.3	50.5±1.9	57.2±3.4	20.9±3.5	26.4±2.8	71.0±0.4	75.9±0.7
SupProto	63.7±4.2	54.6±2.0	59.1±1.9	28.4±3.7	30.2±2.5	70.8±0.3	77.7±1.5
CompMin	64.0±3.3	52.2±1.4	60.6±2.0	24.3±4.1	30.2±2.4	71.8±1.3	77.3±1.0
OrthoProto	**73.9±5.0**	**58.6±2.3**	**66.0±6.1**	**37.0±4.6**	**40.5±7.9**	71.8±1.6	**80.9±1.5**

Table 23: BreastMNIST 15%, batch 256. All metrics ×100, mean±sd over matched seeds. Bold: best mean.

Method	AUROC	ER-AUROC	BAcc	AUPRC$_\pi$	F1$_\pi$	S_{DCA}	$S_{\mathrm{DCA}}^{\mathrm{cal}}$
Logit adj.	97.3±1.0	92.9±1.8	92.1±2.5	94.9±1.5	87.4±1.9	91.4±2.8	93.0±1.7
BCL	94.7±0.8	86.9±1.3	85.8±2.1	89.4±0.8	81.1±2.4	83.7±3.1	89.0±1.2
SupCon	99.4±0.1	97.3±0.4	90.0±9.0	98.5±0.4	94.2±0.3	93.5±3.1	96.4±0.2
KCL	99.4±0.1	97.5±0.2	95.7±0.2	98.6±0.1	92.4±1.6	96.1±0.2	96.6±0.1
PaCo	99.4±0.0	97.4±0.2	95.7±0.2	98.6±0.0	93.7±0.1	95.2±1.0	96.3±0.2
SupMin	99.5±0.1	97.8±0.2	95.8±0.2	98.8±0.1	93.9±0.6	**96.4±0.4**	96.8±0.4
SupProto	99.4±0.2	97.7±0.3	95.5±1.1	98.6±0.3	92.3±1.0	96.0±1.0	97.1±0.5
CompMin	**99.6±0.1**	**98.0±0.2**	**96.4±0.3**	**99.0±0.1**	93.7±2.4	95.9±1.6	**97.2±0.1**
OrthoProto	99.4±0.1	97.7±0.4	95.3±2.0	98.6±0.2	**94.3±0.9**	96.3±0.8	97.1±0.5

Table 24: PneumoniaMNIST native, batch 256. All metrics ×100, mean±sd over matched seeds. Bold: best mean.

Method	AUROC	ER-AUROC	BAcc	AUPRC$_\pi$	F1$_\pi$	S_{DCA}	$S^{\mathrm{cal}}_{\mathrm{DCA}}$
Logit adj.	96.3±0.6	88.9±1.3	85.5±2.1	90.4±1.6	79.4±2.5	82.1±2.7	90.9±0.5
BCL	93.7±0.7	81.8±1.7	83.3±6.5	81.5±2.6	73.4±2.3	80.6±9.0	87.7±0.9
SupCon	99.2±0.1	97.3±0.4	94.9±0.7	98.0±0.3	92.2±1.1	95.8±0.3	96.4±0.3
KCL	99.2±0.2	97.0±0.7	95.8±0.5	97.8±0.5	92.3±0.9	95.6±0.4	96.3±0.4
PaCo	87.6±20.3	83.0±24.7	82.6±22.5	76.2±37.6	73.7±32.1	83.4±21.4	85.9±17.7
SupMin	99.3±0.1	97.0±0.2	95.4±0.5	97.9±0.2	92.6±1.1	95.8±0.2	96.2±0.1
SupProto	99.1±0.1	96.8±0.6	94.1±1.1	97.5±0.4	90.3±2.7	95.0±0.8	96.1±0.6
CompMin	**99.5±0.1**	**97.6±0.4**	95.6±0.5	**98.3±0.3**	**93.0±0.8**	**96.2±0.4**	96.6±0.3
OrthoProto	99.4±0.1	97.6±0.2	**96.0±0.2**	98.1±0.3	93.0±0.4	96.0±0.6	**96.8±0.1**

Table 25: PneumoniaMNIST 20%, batch 256. All metrics ×100, mean±sd over matched seeds. Bold: best mean.

Method	AUROC	ER-AUROC	BAcc	AUPRC$_\pi$	F1$_\pi$	S_{DCA}	$S^{\mathrm{cal}}_{\mathrm{DCA}}$
Logit adj.	95.9±0.6	88.8±0.6	85.1±5.3	87.5±0.9	80.7±0.9	81.8±7.1	90.5±0.9
BCL	95.5±1.0	87.2±2.9	88.3±0.6	84.5±3.3	77.9±3.3	89.3±1.9	90.2±2.1
SupCon	99.1±0.3	96.6±0.3	94.9±0.5	96.8±0.4	90.8±0.4	92.8±2.9	95.8±0.2
KCL	98.8±0.5	96.0±0.9	94.4±1.2	96.2±1.1	89.7±2.0	94.8±1.1	95.3±0.9
PaCo	99.0±0.2	96.2±0.6	95.0±0.4	96.3±0.7	89.8±2.3	94.6±1.4	95.7±0.5
SupMin	98.9±0.2	96.0±0.3	94.3±0.4	96.1±0.3	89.2±1.6	94.0±1.5	95.3±0.5
SupProto	99.1±0.2	96.3±0.8	94.1±0.7	96.3±0.9	90.1±0.6	**94.8±1.0**	95.8±0.8
CompMin	**99.2±0.1**	**96.8±0.2**	92.9±3.7	**97.0±0.2**	**91.2±1.8**	94.6±1.6	96.0±0.2
OrthoProto	99.2±0.4	96.8±1.0	**95.5±1.0**	96.8±1.2	90.0±3.7	94.7±2.5	**96.1±0.7**

Table 26: PneumoniaMNIST 15%, batch 256. All metrics ×100, mean±sd over matched seeds. Bold: best mean.

Method	AUROC	ER-AUROC	BAcc	AUPRC$_\pi$	F1$_\pi$	S_{DCA}	$S^{\mathrm{cal}}_{\mathrm{DCA}}$
Logit adj.	96.8±1.9	91.7±4.3	88.0±5.2	88.0±6.1	77.2±7.1	86.5±7.6	92.5±3.2
BCL	94.3±0.8	83.1±1.7	83.1±1.9	70.7±7.3	63.6±3.2	79.2±2.7	88.3±1.2
SupCon	98.1±0.1	94.7±0.4	93.6±0.4	92.7±0.6	86.8±0.3	90.1±4.8	94.3±0.3
KCL	98.3±0.3	95.4±0.3	93.1±1.7	93.9±0.4	**87.7±2.0**	93.9±0.7	94.5±0.3
PaCo	87.0±19.6	82.3±23.0	82.8±19.5	69.6±42.3	68.8±39.0	82.6±20.3	84.9±17.3
SupMin	98.3±0.0	95.0±0.3	93.5±0.2	93.4±0.1	86.9±3.3	93.6±0.3	94.3±0.2
SupProto	98.6±0.4	95.3±0.8	**94.5±0.5**	93.5±0.9	83.6±7.3	90.4±6.6	94.9±0.6
CompMin	**98.9±0.2**	**96.1±0.8**	93.8±0.9	94.5±1.1	87.4±3.3	91.3±5.7	**95.5±0.8**
OrthoProto	98.8±0.3	96.0±0.8	93.6±1.0	**94.8±1.0**	85.9±4.6	**94.2±0.7**	95.4±0.7

Table 27: PneumoniaMNIST 10%, batch 256. All metrics ×100, mean±sd over matched seeds. Bold: best mean.

Method	AUROC	ER-AUROC	BAcc	AUPRC$_\pi$	F1$_\pi$	S_{DCA}	$S_{\mathrm{DCA}}^{\mathrm{cal}}$
Logit adj.	94.3±1.2	83.7±3.6	76.3±5.1	65.1±7.0	41.8±4.1	72.6±5.1	88.1±1.7
BCL	94.2±1.1	83.2±2.6	83.2±2.8	62.8±8.1	43.4±9.8	80.5±2.1	87.5±1.5
SupCon	97.4±0.3	93.4±1.0	92.6±0.7	85.4±2.2	**80.1±3.7**	92.7±0.9	93.5±0.7
KCL	97.6±0.4	94.1±1.0	**93.2±1.4**	86.9±1.5	79.6±1.0	93.3±0.8	94.0±0.9
PaCo	96.7±1.0	92.7±2.4	91.5±2.8	85.2±4.2	75.6±8.7	89.7±3.9	92.4±2.2
SupMin	97.1±0.6	92.6±1.6	89.3±4.3	84.3±5.4	74.1±5.0	87.1±6.8	92.8±0.9
SupProto	97.5±1.0	93.7±1.2	92.5±0.5	87.9±3.1	78.9±3.4	92.7±0.9	93.1±0.9
CompMin	96.8±0.7	91.5±0.4	90.8±0.8	81.0±0.9	72.6±3.1	89.4±1.0	92.0±0.5
OrthoProto	**98.0±0.4**	**94.8±1.2**	93.1±1.6	**89.8±2.5**	79.0±6.2	**93.6±1.1**	**94.2±1.1**

Table 28: PneumoniaMNIST 5%, batch 256. All metrics $\times 100$, mean±sd over matched seeds. Bold: best mean.

Method	AUROC	ER-AUROC	BAcc	AUPRC$_\pi$	F1$_\pi$	S_{DCA}	$S_{\mathrm{DCA}}^{\mathrm{cal}}$
Logit adj.	94.1±1.0	82.9±4.9	77.3±1.3	38.7±21.9	15.0±4.5	73.9±1.9	88.1±1.3
BCL	92.5±2.6	79.2±6.6	82.9±2.8	34.3±15.9	13.8±3.8	80.4±3.8	86.6±2.9
SupCon	88.1±2.6	68.3±4.2	79.1±2.6	10.7±5.3	8.0±1.2	72.7±5.5	82.0±2.6
KCL	90.0±1.3	72.3±1.2	80.6±1.9	16.5±2.2	9.5±2.0	73.9±4.7	83.3±1.4
PaCo	81.1±4.7	65.3±4.2	69.9±5.4	12.2±5.4	5.2±1.2	66.2±5.9	75.8±3.7
SupMin	92.9±0.8	81.6±3.6	84.1±1.2	29.0±7.5	15.3±3.8	83.4±0.3	86.5±1.8
SupProto	91.4±1.7	77.9±5.0	82.5±1.7	17.5±2.9	9.6±2.2	76.5±10.5	85.2±2.3
CompMin	92.4±1.8	80.9±2.9	83.0±2.9	25.1±11.3	13.1±2.8	82.0±2.3	86.1±2.3
OrthoProto	**96.0±0.5**	**91.1±1.7**	**90.4±1.3**	**64.7±10.4**	**37.6±11.5**	**90.2±0.5**	**91.4±1.3**

Table 29: PneumoniaMNIST 1%, batch 256. All metrics $\times 100$, mean±sd over matched seeds. Bold: best mean.

Method	AUROC	ER-AUROC	BAcc	AUPRC$_\pi$	F1$_\pi$	S_{DCA}	$S_{\mathrm{DCA}}^{\mathrm{cal}}$
Logit adj.	90.9±4.0	75.7±9.4	81.0±3.8	17.8±22.0	6.0±3.8	78.5±3.5	85.0±4.0
BCL	91.6±1.0	80.4±2.3	81.5±0.9	33.0±7.8	6.8±2.8	79.6±5.5	85.0±0.9
SupCon	84.1±1.4	63.3±2.1	70.3±2.5	4.2±1.0	2.8±0.5	69.1±1.8	78.3±1.1
KCL	85.5±1.6	64.5±4.2	70.6±12.6	5.7±0.8	3.4±0.7	67.3±12.5	79.8±1.2
PaCo	79.3±2.9	60.0±1.3	69.9±1.2	3.6±0.2	2.2±0.1	69.0±3.5	74.2±2.0
SupMin	90.1±3.1	75.1±7.2	81.5±3.8	8.8±5.3	5.0±1.3	80.3±2.3	83.9±3.3
SupProto	86.3±3.4	63.6±11.1	79.5±1.1	3.5±2.3	4.2±0.4	78.2±1.9	81.8±1.6
CompMin	90.1±2.0	75.3±4.7	81.0±0.8	8.5±4.2	4.6±0.3	80.9±1.0	84.1±1.5
OrthoProto	**95.0±1.4**	**88.8±2.9**	**88.1±1.3**	**57.0±10.6**	**19.5±11.1**	**87.5±1.5**	**89.8±2.1**

Table 30: PneumoniaMNIST 0.5%, batch 256. All metrics $\times 100$, mean±sd over matched seeds. Bold: best mean.