Auditing Perch V2 in BirdCLEF++ 2026: An eBird Coverage Limit and No Detectable Hard-Selection Headroom

Haochen Shi[1,*]

[1]The Hong Kong Polytechnic University, Hong Kong SAR

Abstract

For BirdCLEF++ 2026 (LifeCLEF Task 2 at CLEF 2026), we audit a pipeline based on the Perch V2 bioacoustic foundation model rather than propose a new competition solution. The task is multi-taxa soundscape recognition in the Pantanal wetlands, scored by macro-averaged AUC across 234 classes. We ask which limits matter once a strong foundation-model anchor is already present: output-vocabulary coverage, hard per-class routing, or focal-to-soundscape transfer. First, the fixed eBird output head of Perch V2 has no label for 31 scored classes—mostly undescribed insect morphospecies—which caps zero-shot macro-AUC at 0.990 under our mapping and scoring conventions. This is a structural output-vocabulary bound; the achieved evaluable macro-AUC of 0.937 is measured only on the mapped/evaluable mask, where remaining errors are recognition errors rather than coverage gaps. Second, hard per-class selection does not provide usable in-domain headroom: across eight anchor–challenger cells, pairing Perch V2 (zero-shot or linear probe) with CNN, PANNs, BEATs, and their combination, no held-out hard-selection layer surpasses the best single model in its cell. One-sided tests exclude positive headroom $\geq +0.01$ in all eight cells and $\geq +0.005$ in seven of eight (Holm-corrected); TOST equivalence at $\delta{=}0.01$ holds for only one cell, so the remaining cells are reported as bounded headroom rather than proven equivalence. Third, calibrated soft fusion gives small in-domain gains, but under focal-to-soundscape shift the Perch V2 zero-shot anchor has the smallest observed drop, and frozen hard-selection or fusion layers do not surpass it. The practical implication is to prioritize non-eBird output paths for unmapped taxa and target-domain adaptation over focal-CV hard routing. We release the audit harness at https://github.com/HaochenSHI66/birdclef2026-audit and cite community components rather than bundling them. Our Bronze result (rank 262/4085) came largely from the attributed community code.

Keywords

BirdCLEF, bioacoustics, foundation models, Perch V2, ensemble headroom, oracle selection, equivalence testing, domain shift, reproducibility

1. Introduction

BirdCLEF++ 2026[1] is the second task within the LifeCLEF track of the CLEF 2026 conference, focusing on multi-taxa soundscape recognition in the Brazilian Pantanal wetlands [1, 2]. The official metric is the macro-averaged ROC-AUC across 234 classes, excluding those with no positive samples; submissions must also satisfy the CPU-only constraints typical of code competitions. As in BirdCLEF+ 2025 [3], the challenge extends beyond leaderboard scores: credible validation, clear provenance, and inspectable failure modes also shape how the results can be interpreted. The STSG BirdCLEF+ 2025 note exemplifies the honest-reporting and reproducibility emphasis we adopt [4].

Our submission earned a Kaggle Bronze medal, ranking 262nd of 4085 teams on the private leaderboard (top 6.4%). We did not build a new state-of-the-art (SOTA) model. Our top-scoring submission relied on a community foundation-embedding stack: Perch V2 embeddings [5], ProtoSSM and PowerOpt community kernels, ecological post-processing, BirdNET gating, and hierarchical taxonomy post-processing. The CNN, PANNs, and BEATs classifiers we trained or adapted within our framework did not improve our final private-leaderboard score; their main role here is to serve as comparison models

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

24130819d@connect.polyu.hk (H. Shi)

0009-0004-1165-1682 (H. Shi)

[1]Official CLEF 2026 materials use both "BirdCLEF+" and "BirdCLEF++" for LifeCLEF Task 2; we write "BirdCLEF++" throughout, and "BirdCLEF+ 2025" for the 2025 predecessor.

for auditing how much additional models contribute once a strong bioacoustic foundation anchor is already present.

This is neither a winning-solution report nor a new modeling-method claim: we reconstruct a recordist-grouped five-fold cross-validation audit, quantify uncertainty, and release the audit harness and outputs needed to reproduce these analyses.[2] Building on the CPU-only ONNX Perch V2 anchor and our trained CNN/PANNs/BEATs models, we examine three practical questions. First, which scored classes cannot map directly into Perch V2's fixed eBird output label space, and what alternatives would be needed for those taxa? Second, given a strong anchor, how much per-class oracle-selection headroom remains over the best single model, and does it persist after a held-out, in-domain hard-selection test on focal-recording cross-validation (focal-CV)? Third, do focal-CV conclusions transfer to the labeled soundscape subset, or does domain shift dominate the smaller in-domain ensemble effects?

Most recent BirdCLEF/LifeCLEF technical reports still present ablation results and leaderboard scores as point estimates. To our knowledge, few attach confidence intervals, permutation nulls, or equivalence tests to their conclusions on ensembling or model selection. Although the field has widely converged on Perch, whether oracle selection under focal-CV has deployable value remains unaudited, and whether coverage—or the shift from focal recordings to soundscapes—is the true limiting factor on performance has not been fully discussed. We therefore ask: once a strong bioacoustic foundation anchor is in place, does the headroom from selecting or fusing additional models hold up after uncertainty quantification and an in-domain hard-selection test? Or is the deployable performance ceiling instead set by coverage and domain shift?

Our contributions (Figure 1) are:

- **An uncertainty-quantified analysis of foundation-embedding BirdCLEF systems**: we give both confidence intervals and permutation nulls for the primary effects. For these effects we provide author-clustered bootstrap intervals, permutation or one-sided exclusion rules, and publicly released artifacts; by contrast, prior working notes have largely reported only point estimates.
- **A foundation-model eBird output-label coverage audit**: of the 234 scored classes, 31 cannot be mapped to Perch's fixed eBird output head; the missing classes fall almost entirely within Insecta, where 25 of the 28 Insecta classes are undescribed morphospecies with no binomial scientific name, no eBird code, and no focal positives. This corroborates the insect/amphibian difficulty highlighted in BirdCLEF+ 2025 [3] and aligns with the broader fixed- versus open-vocabulary recognition problem [6, 7]. We quantify the resulting zero-shot ceiling—0.990 at a scored-metric cost of $\leq$0.010—and identify the alternatives implied by this gap: a non-eBird output head, local sonotype supervision, target labels, pseudo-labels, or external data.
- **A de-biased oracle-headroom audit**: it comprises eight anchor$\times$challenger cells and includes a held-out selection/fusion layer that is conditioned on single-level out-of-fold base predictions but is not fully nested, combined with label-permutation de-biasing [8] and author-clustered bootstrap confidence intervals. The core test is a one-sided exclusion of positive headroom, with a standard error from paired resampling that captures the covariance between the bootstrap and the permutation null. All eight cells exclude headroom $\geq +0.01$, and seven of the eight exclude $\geq +0.005$ (Holm-corrected). Under the stricter δ=0.01 margin, only one cell (ZS+BEATs) passes the two-sided TOST equivalence test; for the other seven we report bounded headroom rather than demonstrated equivalence. We find that, in-domain, hard selection yields no detectable gain over the best single model, consistent with model-selection overfitting theory [9]; re-auditing rare/difficult classes stratified by positive-sample prevalence gives the same null. Calibrated soft fusion is the only strategy with a small positive gain, improving AUC by $\leq$ +0.009 in five of the eight cells. One practically relevant observation is that, on the in-domain anchor task, the fixed eBird output head outperforms the trained linear probe (Perch-ZS 0.936 vs. linear probe 0.920).

[2] Audit harness (fixed folds, cached out-of-fold predictions, scoring, and bootstrap/permutation/TOST code; community components attributed, not redistributed): https://github.com/HaochenSHI66/birdclef2026-audit.

- **Target-domain diagnostics with deployment implications**: the shift from focal recordings to soundscapes is larger than the in-domain ensemble effects. The soundscape macro-AUC is 0.737 over 75 scorable soundscape classes; on the 47 classes shared with focal-CV, the mean per-class AUC drops by 0.039, to 0.889, with an embedding-space MMD^2 of 0.092. Perch is more robust than every challenger checkpoint compared, and a frozen-transfer test shows that neither hard selection nor fusion improves Perch-ZS's target-domain score. Thus focal-CV hard routing is not the main deployment lever in this audit; target-domain adaptation is.

2. Related Work

Foundation bioacoustic and audio embeddings. Perch V2 is the primary pre-trained encoder in our evaluation; paired with a simple downstream head, it yields high-performance, multi-taxa bioacoustic embeddings [5], with most related evaluations using BirdSet [10]. For comparison, we include the general-purpose audio model BEATs (a self-supervised acoustic tokenizer) [11] and PANNs CNN14 (a CNN pre-trained on AudioSet) [12], while treating SurfPerch [13], AVES [14], and the AST [15] model family as relevant encoder architectures. Existing surveys and the BEANS benchmark motivate the "frozen encoder + simple classification head" evaluation paradigm as a meaningful protocol [16, 17]. Research on interpretable prototype networks—such as AudioProtoPNet—also indicates that while Perch performs robustly, it does not dominate across all tasks [18]. We therefore use Perch V2 as a strong frozen-encoder baseline for testing model selection and fusion.

BirdCLEF and focal-to-soundscape adaptation. BirdCLEF 2023 documents the recent competition setting [19]; BirdCLEF 2024 foregrounds focal-to-soundscape adaptation [20]. BirdCLEF+ 2025 is the immediate predecessor to the current setup; it expanded the task to include multi-taxa recognition in CPU-only environments and maintained an emphasis on threshold-free evaluation [3]. The work surrounding DS@GT represents an early attempt to combine off-the-shelf foundation model features with pseudo-labeling techniques for BirdCLEF-style tasks [21, 22]. The STSG BirdCLEF+ 2025 report is a methodological reference for transparency and reproducibility despite its lightweight model not outperforming the baseline [4]. Another system, which secured second place in BirdCLEF+ 2025, addressed the distribution shift between focal recordings and soundscapes through transfer learning and semi-supervised distillation [23]. That study focused on interventions to mitigate the shift, whereas our work focuses on assessing oracle headroom and diagnosing the nature of the shift itself.

Ensembles, complementarity, and calibration. This paper distinguishes between diversity diagnosis and claims regarding ensemble accuracy. Kuncheva and Whitaker systematically discussed metrics such as the Q-statistic and double-fault rate, while also questioning the notion of a simple correspondence between diversity and accuracy [24]. Consequently, we treat complementarity diagnosis as evidence for understanding model behavior, rather than as proof that model fusion will inevitably yield performance gains. Stacked generalization provides the conceptual foundation for learning a calibrated soft combination of model outputs [25]. Regarding calibration, Platt scaling and temperature scaling serve as lightweight baseline methods; existing empirical studies have clarified the conditions under which these scaling techniques are effective versus when they are prone to overfitting [26, 27]. Multi-label bird classification introduces additional constraints: recent research on uncertainty calibration for multi-label bird classifiers advocates for the careful selection of calibration strategies on a per-fold basis and warns against relying on fragile per-class isotonic calibration when classes are scarce [28]. These findings motivated us to adopt "calibrated soft fusion"—specifically, a per-class L_2-regularized logistic regression applied to the base models' logits with signed coefficients. In this context, the method serves as an evaluation tool rather than a proposal for a new ensemble technique.

Equivalence rather than a bare null. Unlike most architecture-focused technical reports, this paper emphasizes interval estimation and equivalence testing. We report author-clustered confidence

Table 1
Selected recent BirdCLEF/LifeCLEF working notes: shared training choices and the one distinguishing component reported by each. Several of these notes report overlapping choices such as log-mel inputs, 5-fold CV, 5 s sliding-window averaged inference, and BCE + AdamW + cosine training; they differ mainly in one distinctive lever. The *Uncert.* column records whether each note reports any quantified uncertainty (CI / significance test / multi-seed dispersion) on its headline results, verified by inspecting each note; "Partial" denotes multi-seed dispersion without an inferential test. "Uncert." counts only explicit inferential uncertainty on headline results (CI / significance / equivalence test); multi-seed point dispersion alone is labeled Partial.

Note	Year/score	Unique angle	Uncert.
Sydorskyi [23]	2025; 0.928 AUC, 2nd	in-domain pretraining	No
Lasseck [33]	2024	pseudo-label loop	No
Witting [34]	2024	shift analysis and TTA	Partial
Hong [35]	2024	progressive KD	No
Miyaguchi [4]	2025	codebook tokens	No
Miyaguchi [22]	2024	frozen BVC features	No
Dmitriev [36]	2024	geographic weighting	Partial
Tan [37]	2025	Mel–MFCC fusion	No
Gokulnath [38]	2025	handcrafted features	No
Sampathkumar [39]	2024	patch reordering	No

intervals for each effect size, use a one-sided exclusion rule for positive oracle headroom above the pre-specified smallest effect size of interest (SESOI), and apply the two one-sided tests (TOST) procedure [29, 30, 31] as a supplementary equivalence check. By reporting both CIs and the results of one-sided exclusion tests, we require that conclusions regarding positive headroom be supported by tests against pre-set thresholds, rather than relying solely on informal point estimates.

Domain-shift diagnosis and concurrent context. Domain shift in bioacoustics can be directly modeled and addressed—for instance, through adversarial robustness training [32]. We instead diagnose focal-to-soundscape differences and remaining hard-selection headroom.

Field landscape and our position. Table 1 summarizes several recent working notes from BirdCLEF/LifeCLEF.

Unlike leaderboard-oriented system reports, this paper isolates oracle selection and fusion headroom under a fixed protocol: we conduct a de-biased, per-class oracle headroom audit in which out-of-fold (OOF) selection and fusion are conditioned on fixed, single-level OOF base predictions—rather than employing a fully nested setup. Statistically, we utilize permutation null distributions, one-sided tests to exclude positive headroom, and supplementary TOST equivalence checks [30, 31]. While stacked generalization and diversity diagnostics are established tools [25, 24]—and the risk of selection overfitting has long been discussed [9]—our contribution lies in quantifying this mechanism and its associated uncertainty to analyze the "foundation anchor" paradigm underpinning competitive BirdCLEF systems, while also supplementing the audit of fixed-output-head coverage [5].

3. Method

3.1. Task, labels, and validation split

This task focuses on passive-acoustic species identification in the Pantanal wetland environment [2] and is subject to code-competition rules: only CPU usage is permitted, and there is no internet access. Submissions must output per-species probabilities for each 5-s window. Evaluation is based on macro-averaged ROC-AUC, excluding classes that lack either a positive or a negative sample (Eq. 1).

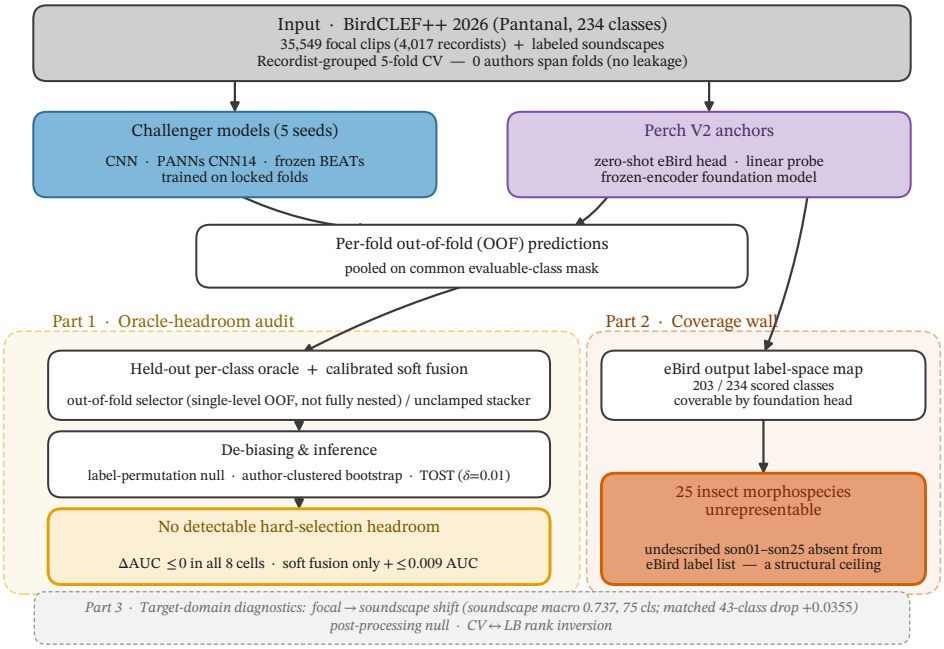

Figure 1: Three-part audit pipeline. Recordist-grouped 5-fold splits feed two Perch V2 anchors (zero-shot, linear probe) and three trained challengers (CNN, PANNs, BEATs), producing out-of-fold predictions on a common evaluable mask. **Part 1 (oracle-headroom audit):** a held-out per-class oracle and calibrated soft fusion, de-biased by a label-permutation null and author-clustered bootstrap with TOST tests, yield no detectable hard-selection headroom across the 8 cells. **Part 2 (coverage wall):** the Perch V2 eBird output head cannot represent 25 undescribed insect morphospecies, a structural ceiling. **Part 3 (target-domain diagnostics):** focal-to-soundscape tests check whether focal-CV conclusions transfer to soundscapes. Quantities appear in the corresponding Results subsections.

The scored label space comprises 234 classes: 162 Aves, 35 Amphibia, 28 Insecta, 8 Mammalia, and 1 Reptilia. The focal training table contains 35,549 clips from 4,017 distinct recordists. We employ a fixed 5-fold StratifiedGroupKFold split, stratifying on `primary_label` and grouping by author. An author-overlap check confirms that no author ID appears in more than one fold. Fold sizes are $\{0{:}5099, 1{:}6847, 2{:}7963, 3{:}8718, 4{:}6922\}$.

Of the 234 scored classes, 206 have focal training recordings, while 28 lack focal examples. Perch V2's zero-shot eBird head covers 203 of the 234 scored classes through a predeclared map: 158 exact matches and 45 scientific-name matches. The 31 unmapped classes comprise 25 undescribed insect morphospecies, named `son01–son25`, and 6 absent taxa: 3 Amphibia, 2 Mammalia, and 1 Reptilia. These undescribed morphospecies have no scientific names and are absent from Perch V2's 14,795 eBird label list, so the fixed eBird head cannot represent them; this reflects an output-vocabulary limit, not an acoustic-transfer claim [5].

3.2. Models

We evaluate two Perch V2 anchors and three supervised challenger families on the same locked folds.

A1 is the Perch V2 zero-shot model: we run the 14,795-dimensional eBird label head and aggregate to the 234 scored classes using the predeclared maximum-aggregation map above. A2 is a supervised Perch V2 linear probe following the frozen-encoder protocol [16]: per fold we extract the 1536-dimensional embedding and fit one-vs-rest logistic regression (`class_weight=balanced`, `max_iter=1000`), with the `StandardScaler` fit on train rows only within each fold.

The challenger families are trained on the same locked folds, with each model using 5 random seeds. S1 is our CNN baseline: a `timm` ImageNet-pretrained backbone with a linear head, taking mel-spectrogram inputs; it is not XC-pretrained. S2 is PANNs CNN14, initialized from AudioSet pretraining [12]. S3 is a frozen BEATs encoder with a trainable head, a general-audio self-supervised family [11].

Each run saves per-example OOF scores, fold IDs, author IDs, class masks, seeds, configs, and raw predictions. The 5 seeds estimate training stochasticity only, not uncertainty arising from splits or fold partitions.

The S1 mel-spectrogram front-end uses a 32 kHz sample rate, 5 s windows, 128 mel bands, a 1024-point FFT, a 320-sample hop, and a 20–16,000 Hz range. The power mel-spectrogram is first converted to a per-clip-max decibel representation, then min–max normalized to $[0, 1]$, and finally bilinearly resized to a square image for the `timm` backbone. This front-end applies to S1 only; S2 (PANNs) uses 64 mel bands over 50–14,000 Hz, while S3 (BEATs) uses its own pipeline.

Except for learning rate and front-end, S1–S3 share identical optimizer settings. All three use AdamW (weight decay 10^{-4}), gradient-norm clipping at 1.0, bfloat16 mixed precision, constant LR, and batch size 32; the base LR is 10^{-3} for S1 (CNN) and S3 (BEATs head) and 5×10^{-4} for S2 (PANNs). Early stopping monitors validation macro-AUC with patience 3, and each of S1–S3 was trained for up to 15 epochs. The reported CNN and PANNs runs use Focal Loss (the default; BCE-with-logits is supported but unused here) with mixup augmentation.

3.3. Compute and reproducibility scope

Reproducibility is scoped to reruns within our environment. Training ran on a 3-card Ascend NPU server using PyTorch and `torch_npu`, with no TensorFlow, JAX, or CUDA stack; Perch V2 embeddings were extracted separately on CPU with ONNX Runtime (`perch_v2_no_dft.onnx`, ≈66 ms per clip). Because the full community TensorFlow/JAX stack could not be reproduced per fold, cross-validation claims are scoped to the Perch V2 CPU-ONNX head and linear probe; the full community bundle serves only as an external leaderboard reference. The CV numbers in the headline tables are tied to a released artifact, command, and seed.

3.4. AUC engine and complementarity analysis

All complementarity analyses use pooled out-of-fold (OOF) predictions on a common mask. With N rows, C label columns ($C = 234$, read as `y.shape[1]`, not hard-coded), K base models, multi-hot targets $y \in \{0, 1\}^{N \times C}$, and OOF probabilities $\hat{p}^{(k)}$, per-class AUC is the Mann–Whitney rank-sum statistic and macro-AUC is its unweighted mean over the mask,

$$\mathrm{AUC}_c = \frac{R_c^+ - n_c^+(n_c^+ + 1)/2}{n_c^+ \, n_c^-}, \qquad \mathrm{mAUC}(\hat{p}; M) = \frac{1}{|M|} \sum_{c \in M} \mathrm{AUC}_c(\hat{p}), \tag{1}$$

where $n_c^+ = \sum_i y_{ic}$, $n_c^- = N - n_c^+$, and R_c^+ uses tie-corrected ranks of $\hat{p}_{\cdot c}$; this implementation was verified against `sklearn` to within $< 1\mathrm{e}{-6}$. A class is evaluable only if it has at least one positive, both labels present, and non-NaN predictions in every compared model,

$$M = \Big\{ c : n_c^+ > 0 \ \wedge \ n_c^- > 0 \ \wedge \ \textstyle\bigwedge_{k=1}^{K} \big(\hat{p}_{\cdot c}^{(k)} \text{ non-NaN}\big) \Big\}. \tag{2}$$

Every headroom-audit hard-selection claim is conditional on focal cross-validation out-of-fold predictions evaluated on M (≈ 194–197 of the 234 scored classes, depending on the anchor pair); it estimates in-domain hard-selection headroom on that evaluable subset and is *not* a 234-class deployment claim. The unmapped and zero-positive classes are by construction outside M.

For each anchor–challenger cell, the candidate set consists of the base models plus their unweighted mean, $\mathcal{C} = \{\hat{p}^{(1)}, \ldots, \hat{p}^{(K)}, \bar{p}\}$ with $\bar{p} = \frac{1}{K} \sum_k \hat{p}^{(k)}$. The outer-fold held selector picks, per class, the candidate with the highest AUC on the other folds (selection set $S_f = \{i : f_i \neq f\}$) and applies it to the evaluation fold ($E_f = \{i : f_i = f\}$),

$$k^\star_{f,c} = \arg\max_{m \in \mathcal{C}} \mathrm{AUC}_c\big(y_{S_f}, \hat{p}^{(m)}_{S_f}\big), \qquad \hat{o}_{ic} = \hat{p}_{ic}^{\,(k^\star_{f_i,c})}, \ \ i \in E_{f_i}, \tag{3}$$

with a fallback to the first base model when no candidate is scorable for (f, c). The same-row apparent oracle is diagnostic only; apparent minus held-out is the optimism gap. Against the best single model $b = \arg\max_k \mathrm{mAUC}(\hat{p}^{(k)}; M_0)$ on the all-model base mask M_0, raw held-out headroom is

$$\Delta_{\mathrm{raw}} = \mathrm{mAUC}(\hat{o}; M) - \mathrm{mAUC}(\hat{p}^{(b)}; M). \tag{4}$$

Uncertainty is estimated using a paired author-clustered bootstrap over the $A = 4{,}017$ authors ($n_{\mathrm{boot}} = 2000$): each replicate resamples authors with replacement, recomputes the shared mask $M^{(r)}$, and yields a headroom estimate $\Delta^{(r)}$, summarized by

$$\bar{\Delta} = \frac{1}{n_{\mathrm{boot}}} \sum_r \Delta^{(r)}, \quad \mathrm{CI}_{95} = [Q_{2.5}, Q_{97.5}], \quad \mathrm{SE} = \mathrm{std}_{\mathrm{ddof}=1}(\Delta^{(r)}), \quad \mathrm{MDE} = 1.96\,\mathrm{SE}, \tag{5}$$

with an approximate minimum detectable effect of $\mathrm{MDE} \approx 0.004$–$0.008$. To remove the selector's intrinsic no-signal selection-variance floor, a label-permutation null [8] permutes labels within author ($n_{\mathrm{null}} = 200$), recomputes the held-out oracle $\hat{o}^{(r)}$, and averages the null (no-signal) headroom; the headline de-biased headroom subtracts this offset from the bootstrap-mean raw headroom $\bar{\Delta}$ of Eq. (5),

$$\hat{\mu}_0 = \frac{1}{n_{\mathrm{null}}} \sum_{r=1}^{n_{\mathrm{null}}} \left[\mathrm{mAUC}(\hat{o}^{(r)}; M^{(r)}) - \mathrm{mAUC}(\hat{p}^{(b)}; M^{(r)})\right], \qquad \Delta_{\mathrm{deb}} = \bar{\Delta} - \hat{\mu}_0, \tag{6}$$

As a secondary reference, we also compute a within-author Gaussian-matched null. The headline standard error $\mathrm{SE}_{\mathrm{deb}}$ is an author-clustered bootstrap ($n_{\mathrm{boot}} = 2000$) in which each resample draws $K = 5$ within-author label permutations; we estimate the de-biased headroom variance as the resample variance of the de-biased statistic minus the within-resample permutation variance divided by K (an ANOVA bias-correction). This captures the bootstrap–null covariance while removing single-permutation Monte-Carlo inflation. Concretely, for each resample r the paired statistic is $S_r = T_r - \bar{N}_r$, where T_r is the raw headroom and $\bar{N}_r$ the within-resample mean over the K=5 permuted-null headrooms; the ANOVA correction targets $\mathrm{SE}^2_{\mathrm{deb}} = \mathrm{Var}_r(T_r - \bar{N}_r) - \mathbb{E}_r[s^2_{N|r}]/K$, and the headline interval is $\Delta_{\mathrm{deb}} \pm 1.96\,\mathrm{SE}_{\mathrm{deb}}$. The best single model is fixed from the pooled OOF comparison; inference is conditional on that observed winner. The Monte-Carlo error of $\mathrm{SE}_{\mathrm{deb}}$, from a delete-1 jackknife over resamples, is $\approx 1{\times}10^{-4}$. This permutation-null de-biaser is an alternative to the order-statistic selection-bias correction of McLatchie and Vehtari [40]. We run equivalence tests by the two one-sided tests (TOST) procedure [29] with margin $\delta = 0.01$ macro-AUC (also 0.005) under a normal approximation [30, 31]. Equivalence holds iff both one-sided p-values fall below 0.05, and the upper test doubles as a one-sided exclusion of positive headroom $\geq \delta$ (these one-sided exclusion p-values are Holm-corrected [41] across the 8 cells):

$$z_{\mathrm{lower}} = \frac{\Delta_{\mathrm{deb}} + \delta}{\mathrm{SE}_{\mathrm{deb}}}, \quad z_{\mathrm{upper}} = \frac{\Delta_{\mathrm{deb}} - \delta}{\mathrm{SE}_{\mathrm{deb}}}, \quad p_{\mathrm{lower}} = 1 - \Phi(z_{\mathrm{lower}}), \quad p_{\mathrm{upper}} = \Phi(z_{\mathrm{upper}}). \tag{7}$$

Complementarity diagnostics include Spearman and Pearson correlation, Kuncheva's Q-statistic, and double-fault rate [24]. Two pipeline controls: force-best-single selection must reproduce $\Delta\mathrm{AUC} = 0$, and the duplicate-anchor control is 0 by construction. These controls are sanity checks, not evidence of model redundancy.

The base out-of-fold columns are single-level global features: each base model saw the outer fold's rows during its own training, so the selection and fusion fits are outer-fold-held conditional on the fixed single-level OOF base predictions, but not fully nested. This inflates the best single model and shrinks apparent residual headroom; we treat a one-seed nested cross-validation re-run (n_{null}=200, n_{boot}=2000), retraining each base model with both the outer and inner fold excluded, as a sensitivity check that estimates how much this bias affects the conclusion (reported in Results and Limitations). To check that the common mask M does not make the null true by construction, we also split each cell's mask into positive-count terciles (Low / Mid / High) and re-run the identical de-biased estimator within each stratum (n_{null}=200, n_{boot}=3000), reusing the released OOF predictions.

3.5. Calibration, fusion, and post-processing

Fusion is a per-class, per-outer-fold L_2-regularized logistic regression on base-model logits (stacked generalization [25]), not temperature/Platt calibration [27, 26] followed by non-negative stacking. With clipped logits $\ell_{ic}^{(k)} = \mathrm{logit}(\mathrm{clip}(\hat{p}_{ic}^{(k)}, \epsilon, 1-\epsilon))$, $\epsilon = 10^{-6}$, the stacker for outer fold f is fit on rows $f_i \neq f$ by `LogisticRegression(C=1.0, max_iter=500)` and applied as

$$\hat{p}_{ic}^{\mathrm{fus}} = \sigma\Big(\beta_{0,c}^{(f_i)} + \sum_{k=1}^{K} \beta_{k,c}^{(f_i)}\, \ell_{ic}^{(k)}\Big), \qquad \sigma(z) = \frac{1}{1+e^{-z}}. \tag{8}$$

The coefficients are sign-unconstrained, with no per-class isotonic calibration on rare classes, because rare-class counts are too small for stable per-class isotonic calibration [28]. Soft-fusion ΔAUC is $\mathrm{mAUC}(\hat{p}^{\mathrm{fus}}; M) - \mathrm{mAUC}(\hat{p}^{(b)}; M)$; summaries are in Table 3.

Post-processing is evaluated as an ablation on the Perch zero-shot out-of-fold base. We test three fixed post-processing settings: genus-level taxonomy smoothing with $\alpha = 0.1$, cross-taxa gating with $\beta = 0.5$, and an ecological spatial prior. The spatial prior uses no evaluation-fold labels and uses global iNaturalist [42] latitude/longitude as a descriptive lower bound; focal recordings are not at Pantanal test sites. The α and β values are not tuned, because tuning on the evaluation mask would be circular. Effects are measured with paired author-clustered bootstrap and Holm/BH-FDR correction over the 3-setting family, and reported in Results.

3.6. Domain-shift and leaderboard validation checks

We run Perch V2 CPU-ONNX on `train_soundscapes_labels.csv` (1,478 labeled multi-label 5-s windows over 66 files) using the same predeclared zero-shot map as in focal validation, and compare per-class ROC-AUC on soundscape windows against focal cross-validation AUC, skipping zero-positive classes and bootstrapping over classes. We also compute unsupervised embedding-shift descriptors: the unbiased RBF kernel maximum mean discrepancy (RBF-MMD^2) [43] with median-heuristic γ, the mean-embedding L_2 gap, and the standard-deviation ratio. For focal embeddings $X = \{x_i\}_{i=1}^{n}$, soundscape embeddings $Y = \{y_j\}_{j=1}^{m}$, and RBF kernel $k(a,b) = \exp(-\gamma\|a-b\|^2)$, the unbiased squared MMD is

$$\widehat{\mathrm{MMD}}_u^2 = \frac{1}{n(n-1)} \sum_{i \neq i'} k(x_i, x_{i'}) + \frac{1}{m(m-1)} \sum_{j \neq j'} k(y_j, y_{j'}) - \frac{2}{nm} \sum_{i,j} k(x_i, y_j), \tag{9}$$

with $\gamma = 1/(\mathrm{median}\{\|z_i - z_j\|^2 : i < j\} + 10^{-12})$ on the pooled sample $Z = [X; Y]$. As a structural bound on zero-shot transfer, the coverage ceiling is the best-case macro-AUC when every mapped evaluable class scores 1.0 and every unmapped evaluable class is pinned at chance 0.5,

$$\mathrm{ceiling}_{\mathrm{zs}} = \frac{1.0 \cdot \big|\{c \in \mathcal{E} : \mathrm{mapped}_c\}\big| + 0.5 \cdot n_{\mathrm{ue}}}{\max(n_{\mathrm{eval}}, 1)}, \qquad \mathrm{drag} = 1 - \mathrm{ceiling}_{\mathrm{zs}}, \tag{10}$$

where $\mathcal{E}$ is the set of evaluable classes ($n_c^+ > 0$, two-class), $n_{\mathrm{eval}} = |\mathcal{E}|$, and n_{ue} counts evaluable classes with no Perch eBird index; this is an output-vocabulary bound, not an acoustic-transfer claim.

For the Perch-versus-challenger robustness diagnostic, we retrain each of the three challenger families (CNN, PANNs, BEATs) on folds 0–3, hold out fold 4, save each checkpoint (e.g. `cnnrg_fold4_seed42.pt`), and re-infer the same 66-file soundscape set. This is a single-checkpoint-per-family diagnostic on the 43 shared classes, not a replacement for the 5-seed/5-fold focal-CV ensemble. For the pseudo-labeling audit, we train two otherwise identical CNNs: focal-only and focal plus Perch-ZS top-1 pseudo-labeled soundscape windows. Pseudo-labels are treated as model outputs, not ground truth.

Finally, we report CV-to-private-LB behavior only for the 3 reproducible validation configurations recorded before late submissions and submitted once after the deadline: Perch-ZS, a single-checkpoint CNN, and their untuned equal-weight fusion. This is a target-domain validation check with $n = 3$, not a rank-correlation claim and not evidence about the separate community competition stack.

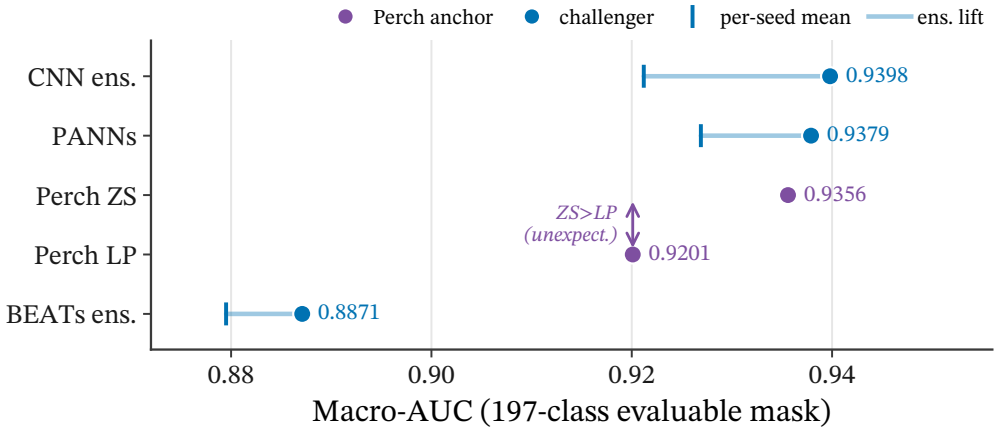

Figure 2: Pooled macro-AUC on the common 197-class evaluable mask (recordist-grouped 5-fold CV, pooled OOF) for the two Perch anchors (purple) and three trained challengers (blue), sorted by AUC. For the trained systems the tick marks the per-seed mean and the segment runs to the 5-seed ensemble dot ($+0.013$–0.019 ensembling lift; per-seed std ≤ 0.0017). Perch zero-shot (0.936) exceeds its own linear probe (0.920), so the probe anchor is "strong" only relatively.

4. Results

4.1. Single-model baselines

On the common 197-class mask, the trained CNN ensemble achieved the highest pooled macro-AUC (0.9398), followed by PANNs CNN14 (0.9379) and the Perch V2 zero-shot model (0.9356); the Perch linear probe (0.9201) and the frozen BEATs ensemble (0.8871) were lower (Figure 2). The macro-AUC values reported below use different masks—the common 197-class mask here, 0.937 on an evaluable subset, and 0.9368 on a 202-class post-processing baseline—so they cannot be compared directly. Variation across random seeds during training was minimal (seed std ≤ 0.0017); seed ensembling improved these challengers by approximately $+0.013$ to $+0.019$ AUC over their per-seed means.

The fixed eBird head beats a trained linear probe (ZS > LP). On the in-domain anchor, the Perch V2 zero-shot model outperformed its *own* supervised linear probe (0.9356 vs. 0.9201; a $+0.0155$ ZS−LP margin), showing that a supervised probe is not necessarily the optimal adaptation. The linear-probe baseline therefore serves only as a relative reference point for the oracle-headroom analysis.

4.2. Coverage wall of the foundation label space

The foundation head leaves 31 of the 234 scored classes unscorable; this gap stems from constraints in the eBird output vocabulary rather than an inability of the encoder to transfer acoustic features. Table 2 gives the mapping for all 162 Aves classes; among the others, 25 of 28 Insecta, 3 of 35 Amphibia, 2 of 8 Mammalia, and the single Reptilia class remain unmapped. The unmapped species are itemized below the table. These 25 unmapped insect classes are morphospecies that have not yet been formally described (`47158son01–25`); they carry only placeholder sonotype labels rather than binomial scientific names, and no focal positives. They therefore fall outside both Perch's fixed eBird output space and the focal supervision signal. A further 6 classes possess valid binomial names but are also absent from the eBird vocabulary. The encoder may carry usable signal with no eBird output head to express it, so supervised challengers remain essential wherever focal data exist.

If all mapped, evaluable classes scored perfectly while all unmapped, evaluable classes performed at chance (0.5), the theoretical upper bound for the zero-shot macro-AUC would be 0.990. Because only 4 of the 31 unmapped classes are AUC-evaluable (10 focal positives in total), the actual score loss attributable to the coverage gap is no more than 0.010; therefore, the 0.990 upper bound should not be compared directly with the achieved evaluable macro-AUC of 0.937. Supervised challengers are trained

Table 2
Coverage × taxon analysis of the 234 scored classes. "Mapped" = representable in Perch V2's fixed eBird output vocabulary (exact eBird-code matches + scientific-name fallbacks); "Unm." = unmapped. "Median pos." is the per-class focal train-recording count (median [min–max]). "n_{AUC}" counts classes with a computable per-class AUC (focal-CV / zero-shot); zero-shot AUC is null for unmapped classes by construction. Per-class AUC is mean (median). The coverage wall falls structurally on *Insecta* (25/28 unmapped, median 0 focal positives—undescribed sonotype morphospecies), while all 162 *Aves* classes map and transfer normally ($\mathrm{AUC} \approx 0.94$).

Taxon	Classes	Exact	Fallb.	Unm.	Median pos.	Focal-CV	Zero-shot
Aves	162	158	4	0	199.5 [3–1105]	0.937 (0.941)	0.935 (0.939)
Amphibia	35	0	32	3	6 [0–63]	0.964 (0.984)	0.953 (0.997)
Insecta	28	0	3	25	0 [0–181]	0.987 (0.987)	0.969 (0.984)
Mammalia	8	0	6	2	7.5 [1–45]	0.853 (0.937)	0.888 (0.879)
Reptilia	1	0	0	1	1 [1–1]	–	–
Total	234	158	45	31	123 [0–1105]	0.938 (0.947)	0.937 (0.943)

directly on these classes, so the unmapped set constrains zero-shot retrieval rather than the supervised models themselves.

The coverage×taxon breakdown (Table 2) indicates that the gap is primarily concentrated in *Insecta* (25 of the 28 unmapped classes); in contrast, mapped classes containing positives typically achieve higher per-class AUC in focal cross-validation (focal CV). Two asymmetries in the mapped non-bird taxa are worth commenting on. First, *Amphibia* maps well (32 of 35 classes) and its evaluable per-class AUC is comparable to Aves, indicating that once an eBird handle exists the encoder transfers as well to anurans as to birds. Second, *Mammalia* both has fewer classes (6 of 8 mapped) and a lower mean AUC that is dominated by a handful of sparsely represented classes; the small denominator makes any single hard or noisy class visibly move the average. Read together, the pattern suggests that coverage is the binding limit for Insecta, whereas for Mammalia coverage is compounded by both a small evaluable pool and per-class recognition difficulty, rather than a simple binary of "covered = solved."

Addressing the unmapped species requires changing the output pathway rather than only reranking Perch's eBird scores. The intervention depends on which evidence is missing. Named taxa that are absent from eBird but have focal positives can be handled first with a local supervised head or a non-eBird taxonomic head. Sonotype labels with focal positives need local sonotype supervision rather than eBird aggregation. The 25 zero-positive sonotypes require a different route: target labels, carefully validated target-domain pseudo-labels, or external insect/amphibian recordings with compatible taxonomy. A supervised head trained only on the current focal data cannot solve those zero-positive classes by itself.

n_{AUC} (focal-CV / zero-shot) per taxon: Aves 162/162, Amphibia 28/31, Insecta 2/3, Mammalia 7/6, Reptilia 0/0; total 199/202. Focal-CV macro over 199 evaluable classes $= 0.938$ (cf. pooled CNN-ensemble 0.940).

Unmapped detail (all 31 classes outside Perch's eBird output vocabulary). The 25 unmapped Insecta are undescribed morphospecies (`47158son01–25`; no binomial, no eBird code, 0 focal positives). The other 6 have valid binomials but are still absent from Perch's output vocabulary (focal positives in parentheses): *Caiman yacare* (Reptilia, 1), *Adenomera guarani* (Amphibia, 0), *Lysapsus limellum* (Amphibia, 5), *Chiasmocleis mehelyi* (Amphibia, 0), *Sapajus cay* (Mammalia, 1), *Mico melanurus* (Mammalia, 3).

Takeaway. The coverage gap is a fixed-output-vocabulary limit, not an acoustic-transfer failure: it is concentrated almost entirely on undescribed insect morphospecies (25 sonotype labels with no Linnaean binomial, no eBird code, and 0 focal positives), whereas every in-vocabulary class with positives—across all taxa—transfers at high AUC.

Table 3
Headline oracle-headroom audit across 8 anchor×challenger cells. De-biased ΔAUC = held-out per-class hard-selection oracle − best single, recentred on a label-permutation null (n_{boot}=2000, n_{null}=200). De-biased 95% CIs use the paired-resampling standard error (raw and permuted-null headroom recomputed on the same author-bootstrap resample, capturing their covariance). A one-sided test rejects positive headroom $\geq +0.005$ in 7 of 8 cells and $\geq +0.01$ in 8 of 8 (Holm-corrected); two-sided TOST equivalence at δ=0.01 holds in 1 of 8 (ZS+BEATs, amber), so the other seven are bounded headroom (TOST uses two one-sided 0.05 tests, a ≈90% interval, by design). Soft-fusion ΔAUC = calibrated held-out fusion − best single; highlight marks a soft-fusion-exploitable gain (CI $\not\ni 0$, *).

Anchor	Chall.	Best	Hard Δ	Hard CI	TOST	Fusion Δ	Fusion CI
LP	CNN	0.9396	-0.0061	$[-0.0159, +0.0036]$	no	$+0.0011$	$[-0.0014, +0.0038]$
LP	PANN	0.9365	-0.0023	$[-0.0120, +0.0074]$	no	$+0.0043$	$[+0.0013, +0.0078]$*
LP	BEATs	0.9202	-0.0183	$[-0.0257, -0.0109]$	no	$+0.0030$	$[-0.0018, +0.0083]$
LP	all3	0.9396	-0.0087	$[-0.0189, +0.0016]$	no	$+0.0045$	$[+0.0017, +0.0076]$*
ZS	CNN	0.9412	-0.0234	$[-0.0308, -0.0161]$	no	$+0.0069$	$[+0.0042, +0.0095]$*
ZS	PANN	0.9379	-0.0132	$[-0.0196, -0.0068]$	no	$+0.0092$	$[+0.0050, +0.0139]$*
ZS	BEATs	0.9355	-0.0031	$[-0.0103, +0.0041]$	yes	-0.0000	$[-0.0032, +0.0033]$
ZS	all3	0.9412	-0.0141	$[-0.0213, -0.0069]$	no	$+0.0090$	$[+0.0067, +0.0114]$*

4.3. Per-class hard selection and soft fusion

Based on focal-CV out-of-fold predictions and the common evaluable-class mask (196 classes for the linear-probe cells, 194 for the zero-shot cells), there is no detectable per-class hard-selection headroom over the best single model: across all eight cells the de-biased ΔAUC is consistently ≤ 0, and positive in-domain gains (up to $+0.009$ AUC) appear only with calibrated soft fusion. The headline decision rule is a one-sided exclusion of positive headroom $\geq +0.005$ (which holds in 7 of 8 cells after Holm correction) and $\geq +0.01$ (which holds in all 8 cells); as a stricter secondary check, the two-sided TOST equivalence test at $\delta = 0.01$ [29, 30, 31] holds in only one cell (ZS+BEATs), so for the remaining seven we report bounded headroom rather than demonstrated equivalence. This is consistent with selection-overfitting theory [9]; evaluations rest on a label-permutation null [8] and an author-clustered bootstrap ($n_{\text{boot}} = 2000$, $n_{\text{null}} = 200$).

Table 3 reports the de-biased headroom (ranging from -0.0023 to -0.0234 AUC across the 8 cells). Uncertainty is estimated by an author-clustered bootstrap with K=5 within-author permutations per resample (using an ANOVA bias-correction; see Methods), capturing the bootstrap–null covariance while avoiding the variance inflation of single-permutation Monte-Carlo simulation; per-cell standard errors (SE) range from 0.0033 to 0.0052 (Monte-Carlo error $\approx 1 \times 10^{-4}$). One-sided tests show that, across the 8 cells, 7 exclude positive headroom $\geq +0.005$ (all but LP+PANN; maximum Holm-adjusted $p = 0.038$), and all 8 exclude $\geq +0.01$ (maximum Holm-adjusted $p = 0.0066$). This is a power-bounded null test for per-class hard selection; it does not amount to a denial of complementarity, and—except for ZS+BEATs—it does not demonstrate equivalence at δ=0.01.

As a limited LP-only subset sensitivity check, we ran a single-seed nested cross-validation re-run (n_{null}=200, n_{boot}=2000), retraining the base models after excluding both outer- and inner-fold data, over four LP combinations (CNN, PANN, CNN+PANN, all3): the de-biased headroom did not differ significantly from zero, and every 95% CI included zero (LP+CNN -0.0048, CI $[-0.012, 0.002]$; LP+PANN -0.0049, CI $[-0.013, 0.002]$; LP+CNN+PANN $+0.0001$, CI $[-0.006, 0.007]$; LP+all3 -0.0061, CI $[-0.014, 0.002]$). This subset does not cover LP+BEATs or any zero-shot cell, so it does not directly validate those cells or the full eight-cell headline. The two combined-challenger cells sit essentially at zero, consistent with the headline. Within this subset the estimates track the global-OOF estimates closely (maximum absolute difference 0.0068), so the LP-only finding is not an artifact of single-level global OOF base features. Across 20 repeated author-fold split seeds, no cell exhibited a de-biased headroom above 0.01 under any seed (maximum $+0.0002$ across all seeds). A 4×4 selector-shrinkage grid over selection margin and minimum-positive thresholds never produced held-out headroom above ≈ 0.004

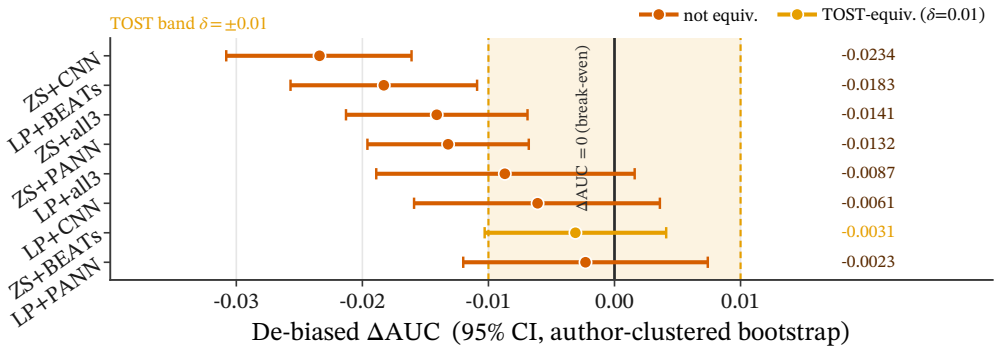

Figure 3: Headline oracle-headroom audit: de-biased per-class hard-selection ΔAUC with paired-resampling 95% CIs, for all 8 anchor×challenger cells (ZS = Perch zero-shot, LP = linear probe). Every point estimate is ≤ 0; amber band marks the $\delta = \pm 0.01$ TOST reference zone; only ZS+BEATs is TOST-equivalent at δ=0.01, the other seven cells are reported as bounded headroom. The best single model varies by cell, so "no headroom" means the held-out selector did not beat that cell's own best single model.

in any configuration. All cross-checks passed: the AUC scorer matched `sklearn` exactly (maximum absolute difference 0); the split-leakage report showed 0 authors and 0 files spanning folds, with no duplicate files; and a leave-one-author influence analysis left every cell's verdict unchanged (de-biased range $[-0.0074, -0.0034]$, no sign flips). A leave-one-file analysis was not performed; leave-one-author is a coarser influence check.

Calibrated fusion yielded AUC gains from $+0.0011$ to $+0.0092$; confidence intervals excluded zero in 5 of the 8 cells (LP+PANN, LP+all3, ZS+CNN, ZS+PANN, ZS+all3), with the largest gains in the zero-shot cells. Even where held-out hard selection showed no detectable headroom, calibrated soft fusion could still exploit inter-model complementarity [25].

The selection-overfitting pattern [9] was most pronounced on the zero-shot anchor: apparent oracle scores were around 0.94–0.95 against held-out selector scores of 0.905–0.920 (optimism gaps 0.020–0.046), and raw held-out hard selection actually hurt the anchor (pre-de-biasing ΔAUC of -0.027 to -0.036). Diversity diagnostics are consistent with this but do not explain it: in the LP cells, Kuncheva Q was 0.981–0.997 with a double-fault rate of about 0.0023, indicating highly similar error outcomes [24]; in the ZS cells, Q was negative (-0.66 to -0.79) with lower double-fault (0.0003–0.0006), yet this apparent diversity did not translate into hard-selection gains.

Not an artifact of challenger weakness. The lack of positive headroom cannot be attributed simply to the weakest challenger. Even for the strong challengers CNN and PANNs (standalone macro-AUC 0.9398 and 0.9379), the de-biased headroom remains negative under either Perch V2 anchor (CNN -0.0234 ZS, -0.0061 LP; PANNs -0.0132 ZS, -0.0023 LP). BEATs is anchor-dependent: closest to zero under the zero-shot anchor (-0.0031) but most negative under the linear-probe anchor (-0.0183). BEATs alone therefore cannot settle the question, and its linear-probe result remains compatible with model weakness.

Not an artifact of dropping hard classes. The common mask M may exclude rare or hard classes where complementarity is most likely. To examine this within the evaluable mask, we split each cell's mask into positive-count terciles and re-ran the same de-biased estimator within each stratum (n_{null}=200, n_{boot}=3000; Table 4). In the Low (rare/hard) stratum, *no* cell exhibited positive-detectable headroom. Across all 24 cell×stratum combinations, the only positive-detectable case was ZS+BEATs in the High stratum (Δ_{deb}=$+0.0097$, CI $[+0.0084, +0.0112]$), showing that the procedure can register a positive signal under the same rule, though this is not a broad guarantee of statistical power.

* Sole positive-detectable stratum across all 24 (CI $\not\ni 0$, sign > 0). All other 23 strata have $\Delta_{\text{deb}} \leq 0$ or a CI that includes 0; in the Low stratum specifically, no cell is positive-detectable.

Figures 3–5 visualize these results.

Table 4
Prevalence-stratified de-biased hard-selection headroom. The per-cell common evaluable mask M is split into positive-count terciles (Low / Mid / High); each stratum is re-audited with the same held-out nested oracle − best-single estimator, label-permutation de-biasing, and author-clustered bootstrap CI restricted to the stratum's classes (n_{null}=200, n_{boot}=3000). "Med. pos." is the median per-class focal-positive count in the stratum. This addresses the "null-by-construction" objection that M deletes the hard/rare classes where headroom would live: in the *Low* (rare/hard) stratum *no* cell is positive-detectable. The *single* positive-detectable cell across all 24 strata is ZS+BEATs High ($+0.0097$, CI $[+0.0084, +0.0112]$), showing the procedure can register a positive signal under the same rule (not a general power guarantee).

Stratum	Median pos.	Cell	Δ_{deb}	95% CI	n
Low	∼31	LP+CNN	-0.0036	$[-.019, +.010]$	66
Low	∼31	LP+PANN	$+0.0038$	$[-.010, +.016]$	66
Low	∼31	LP+BEATs	-0.0357	$[-.045, -.027]$	66
Low	∼31	LP+all3	-0.0051	$[-.020, +.008]$	66
Low	∼31	ZS+CNN	-0.0219	$[-.036, -.009]$	65
Low	∼31	ZS+PANN	-0.0090	$[-.019, +.001]$	65
Low	∼31	ZS+BEATs	-0.0160	$[-.027, -.005]$	65
Low	∼31	ZS+all3	-0.0157	$[-.029, -.003]$	65
Mid	∼156	LP+CNN	-0.0051	$[-.011, +.000]$	65
Mid	∼156	LP+PANN	-0.0046	$[-.011, +.002]$	65
Mid	∼156	LP+BEATs	-0.0116	$[-.016, -.008]$	65
Mid	∼156	LP+all3	-0.0084	$[-.013, -.005]$	65
Mid	∼156	ZS+CNN	-0.0314	$[-.042, -.020]$	64
Mid	∼156	ZS+PANN	-0.0064	$[-.013, +.003]$	64
Mid	∼156	ZS+BEATs	-0.0030	$[-.007, +.001]$	64
Mid	∼156	ZS+all3	-0.0128	$[-.020, -.003]$	64
High	∼449	LP+CNN	-0.0012	$[-.004, +.001]$	65
High	∼449	LP+PANN	-0.0027	$[-.006, +.001]$	65
High	∼449	LP+BEATs	-0.0095	$[-.012, -.007]$	65
High	∼449	LP+all3	-0.0053	$[-.007, -.004]$	65
High	∼449	ZS+CNN	-0.0146	$[-.019, -.009]$	65
High	∼449	ZS+PANN	-0.0231	$[-.031, -.013]$	65
High	∼449	ZS+BEATs	$+0.0097$	$[+.008, +.011]^*$	65
High	∼449	ZS+all3	-0.0115	$[-.015, -.007]$	65

4.4. Post-processing ablations

On the Perch V2 zero-shot CV baseline (0.9368 over 202 classes), none of the three post-processing knobs helped: taxonomy smoothing (α=0.1) was a null (ΔAUC $+0.0008$, CI $[-0.0015, +0.0028]$), cross-taxa gating (β=0.5) significantly hurt (-0.0032, CI $[-0.0050, -0.0023]$), and the ecological spatial prior (global iNaturalist) hurt strongly (-0.0306, CI $[-0.0382, -0.0239]$); effects use a paired author-clustered bootstrap with Holm correction.

The spatial prior used global iNaturalist coordinates rather than Pantanal test-site coordinates, so its effect is the cost of a mismatched spatial prior, not a bound on a correctly localised site prior. On the separate community leaderboard stack, post-processing produced only a small private-LB change on the bronze stack (see Table 5); these observations are not interchangeable.

4.5. Target-domain diagnostics

The five target-domain checks share a common limitation: they all rely on the same small labeled soundscape set (66 files, 1,478 windows; 43–47 common classes depending on the mask) and are therefore *power-bounded*—underpowered for small effects, though adequate to bound larger target-domain levers. The largest observed effect is the focal-to-soundscape drop, and no focal-learned combiner beat the

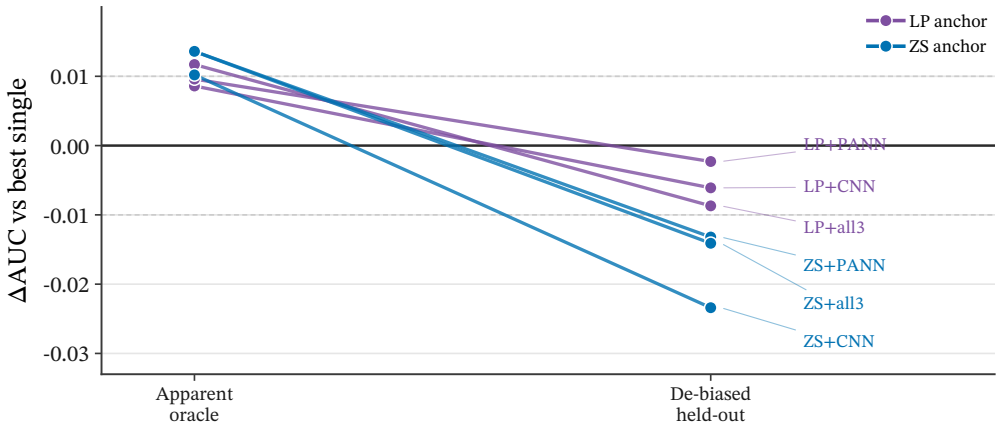

Figure 4: Selection-overfit anatomy, tracing six cells (BEATs omitted for readability) from apparent oracle ΔAUC to de-biased held-out ΔAUC; the held-out estimates collapse below zero. Purple denotes the LP anchor and blue denotes the ZS anchor.

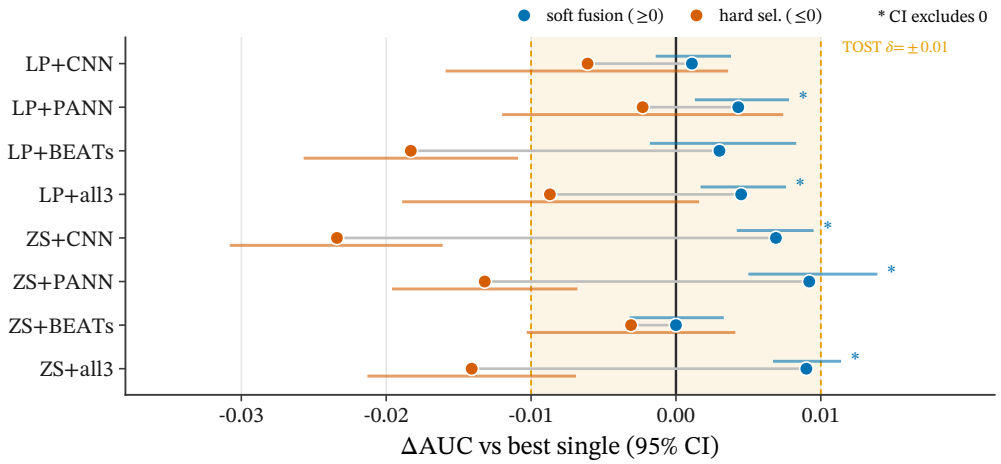

Figure 5: Calibrated soft-fusion ΔAUC (blue, * CI excludes 0 in 5/8 cells) compared with de-biased hard-selection ΔAUC (red).

Perch V2 zero-shot model on the target.

Focal-to-soundscape shift. The Perch V2 zero-shot model [5] dropped substantially on the soundscape set. Over the 75 scorable soundscape classes its macro-AUC is 0.737; over the 47 classes shared with focal CV the mean per-class AUC drop is $+0.0393$ (class-bootstrap 95% CI $[0.0010, 0.0835]$), to 0.889. The drop was class-specific: 20 of the 47 classes dropped, worst for redjun ($0.899 \rightarrow 0.481$) and strher2 ($0.936 \rightarrow 0.534$), while others were barely affected. Unsupervised embedding diagnostics agree: unbiased RBF-MMD$^2 = 0.0917$ [43], the mean-embedding L2 gap is 2.03, and the standard-deviation ratio is 0.95. This shift is large relative to the focal-CV ensemble effects above (Figure 6).

Robustness gap: Perch vs. CNN, PANNs, and BEATs. Each challenger family was trained on folds 0–3, held out on fold 4, and re-inferred on the same 66-file soundscape set. Perch V2 zero-shot dropped least (0.0355 on this matched 43-class fold-4 mask); each challenger dropped far more (CNN 0.1984, PANNs 0.1847, BEATs 0.1815), so on the matched 43-class mask the challenger-minus-Perch drop gap was positive for all three ($+0.1629/+0.1492/+0.1460$; Figure 7). Part of this apparent robustness is structural: the zero-shot anchor was never fit to the focal data, so it has no focal-specific signal to lose under the shift. Each checkpoint here is a single fold-4-held model, not the 5-seed/5-fold ensemble used in the focal-CV audit.

The practical consequence is that focal-CV ranking alone is a weak deployment criterion for this

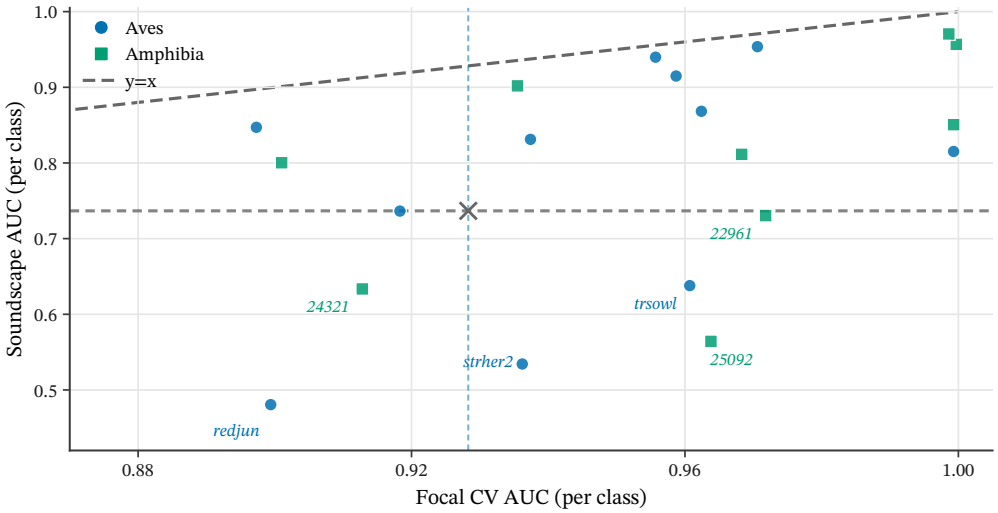

Figure 6: Focal-CV AUC versus soundscape AUC for the 20 most-affected of 47 common scorable classes under Perch V2 zero-shot (circle=Aves, square=Amphibia); *redjun* drops $0.899 \rightarrow 0.481$, the 47-class shared mask drops from 0.928 to 0.889, the 75-class soundscape macro-AUC is 0.737, and RBF-MMD^2=0.092.

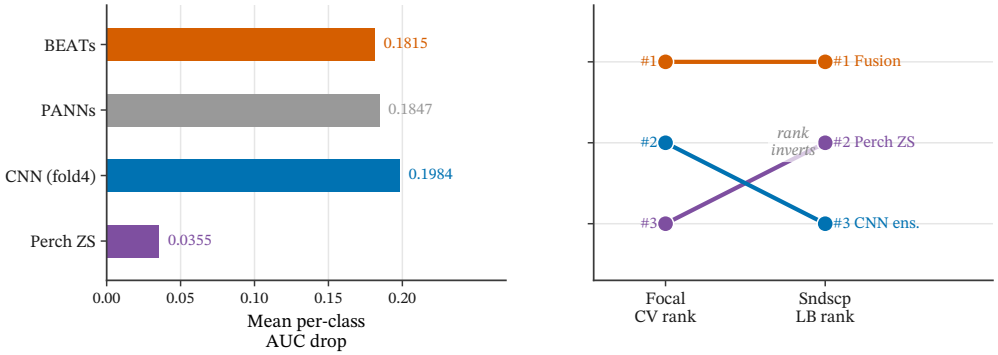

Figure 7: Robustness gap and CV–LB rank inversion. Perch ZS has the smallest mean per-class focal→soundscape AUC drop on the matched 43-class fold-4 mask (0.0355 versus CNN 0.198, PANNs 0.185, BEATs 0.182), and Perch ZS and the CNN ensemble swap rank between focal CV and the soundscape LB validation track (n=3, qualitative).

task. A model or combiner that wins on focal recordings can lose once evaluated on soundscapes, so target-domain validation or adaptation should precede any hard-routing decision.

Validation-track CV–LB gap and rank inversion. We obtained one late, pre-registered, single-submit private-LB score per author-reproducible validation configuration (Table 5); these are not the community competition stack and did not affect the official rank. They reproduce a rank inversion on the hidden test: focal CV ranks CNN above Perch (0.940 vs. 0.936), while the shifted soundscape LB ranks Perch above CNN (0.8419 vs. 0.8097). Equal-weight fusion scored highest here but without CIs; we do not read these n=3 points as evidence that fusion grows under shift (see the frozen-transfer test below).

Frozen focal-to-soundscape transfer. To test whether a focal-learned combiner transfers, we froze each focal-learned layer and applied it unchanged to the labeled soundscape set (45 classes, file-clustered bootstrap $n_{\text{boot}} = 2000$). The Perch zero-shot anchor was the best single model on the target (macro-AUC 0.891). Frozen hard selection *hurt* (ΔAUC -0.078, CI $[-0.112, -0.039]$) and frozen soft fusion did not help (-0.012, CI $[-0.030, +0.007]$). No learned combiner beat the zero-shot anchor under shift: the small in-domain soft-fusion gain does not become a detectable target-domain gain once frozen.

Pseudo-labeling audit. We tested soundscape pseudo-labeling—a dominant lever in recent BirdCLEF systems—as a clean target-domain intervention. Two identically-configured CNNs (focal-only baseline vs. focal-plus-PL, the latter trained with Perch-ZS top-1 pseudo-labels over ≈9000 clips, 148 classes) scored 0.8031 and 0.8046 macro-AUC on the 45-common-class soundscape set, for ΔAUC $+0.0015$ (file-clustered 95% CI $[-0.030, +0.039]$; focal-CV effect neutral at -0.0015). Under the shared target-domain caveat above, this is no detectable headroom rather than a refutation of pseudo-labeling: the zero-shot teacher was weak and the run used a single seed.

4.6. Provenance note: official competition leaderboard track

Separately from the reproducible CV audit and the late validation submissions above, we record the official Kaggle private leaderboard result. The Bronze-attributed competition stack (NFNet-first plus EoS8 50/50, `53225430`, 2026-05-31) scored 0.95044 public and 0.94221 private LB, rank 262 of 4085 (top 6.4%); the best community EoS8 private score was 0.94247. On the attributed stack, author from-scratch acoustic sidecars produced zero private-LB movement (engineered as budget-clamped rank corrections of ≈0.006 maximum). A late post-deadline rerun (`53466774`, 2026-06-08) scored 0.94224 private and confirmed the stack end-to-end without changing the official rank.

5. Discussion

The core conclusion is negative: with a robust bioacoustic foundation anchor, our experiments do not support hard per-class routing atop Perch V2. One-sided tests revealed that all eight cells rejected the hypothesis of a positive headroom $\geq +0.01$, with seven of the eight rejecting the hypothesis of $\geq +0.005$ (Holm-corrected); only calibrated soft fusion [25] recovered some complementarity, though the resulting in-domain gains were minimal. Consequently, we characterize this result as bounded headroom established via one-sided exclusion, rather than demonstrating equivalence at δ=0.01 (a condition met in only one cell, ZS+BEATs) [31]. Error diversity did not translate into routing gains on the held-out set. Negative Kuncheva Q values in some zero-shot cells did not translate into held-out selection gains [24], because diverse error patterns need not have aligned score scales or confidence levels. Before relying on such diagnostic results, validation on held-out prediction data is essential [27, 28].

Practical takeaways. First, focal-CV hard routing should not be the default next layer unless it also wins on held-out soundscape-like data. Second, unmapped taxa need an output path outside Perch's fixed eBird head, and zero-positive sonotypes require new positive evidence rather than only a new classifier. Third, leaderboard claims should be separated from author-reproducible validation tracks, because the community competition stack and the audit reruns answer different questions.

The Perch/eBird pipeline employs a fixed output vocabulary, and our coverage audit quantified the implications of this constraint [5]. This result highlights the performance gap between fixed-vocabulary and open-vocabulary recognition [6, 7] and aligns with the challenges in identifying Insecta and Amphibia noted in the BirdCLEF+ 2025 overview [3]. For "long-tail" classes outside the base classification taxonomy, supervised challengers remain necessary where focal data exist. For the 25 "zero-positive-sample" undescribed sonotypes, however, neither the fixed eBird output head nor a supervised head trained on existing focal data can assign scores. Bridging this gap requires non-eBird output pathways—for named non-eBird taxa, a supervised classifier trained on a taxonomic vocabulary that includes the missing families (or on placeholder sonotype labels where focal positives exist); and for the 25 zero-positive sonotypes, retrieval against externally curated insect-recording collections or open-vocabulary embedding matching, since no fixed head trained only on current focal data can score them—together with future positive evidence (target labels, target-domain pseudo-labels, or external recordings with compatible taxonomy), rather than relying solely on supervised learning over the existing focal set. Because zero-shot Perch V2 outperforms its linear probe, the base output head may

already be the stronger baseline. Gains from post-processing and pseudo-labeling follow a similar pattern: before being regarded as genuine performance improvements, they require re-testing on clean data within the target domain.

In this audit, the largest observed challenger loss is the domain-shift gap: for the tested challenger checkpoints on the matched 43-class fold-4 mask, the focal-to-soundscape drop (≈ 0.19) is far larger than the Perch-ZS drop on the same mask (0.0355), exceeds every cross-validation (CV) ensembling and post-processing effect measured here, and helps explain the CV–leaderboard discrepancy. The rank inversion between the CNN ensemble and Perch-ZS makes the implication concrete: focal-CV rank is useful for debugging, but it should not be the sole criterion for choosing a deployable soundscape model. Two operational consequences follow. First, model-selection procedures that use focal-CV as their only tie-breaker can systematically favor overfit challengers over robust foundation embeddings, so at least one soundscape-domain validation split should be reserved before hard-routing decisions are locked. Second, that a zero-shot anchor—never fit to focal data—beats every focal-trained challenger under shift is evidence that part of the accuracy extracted from focal-CV training is focal-recording-specific signal (recordist, background, microphone) rather than transferable class discrimination; this makes "robustness under shift" a first-class selection criterion, not a post-hoc sanity check. A focal-learned selector or fusion layer should be re-tested on soundscape-like data before being trusted. In practice, soundscape adaptation should take priority over hard routing, with attention to temporal context, background noise, site calibration, and overlapping vocalizations. Recent research on intervention strategies has directly addressed these issues [23, 32]; the contribution of this study lies in providing a diagnosis and pinpointing the stages where performance loss is most severe.

6. Limitations

Statistical limitations constrain the estimation of "headroom." The base-model "out-of-fold" (OOF) predictions are single-level global OOF features rather than being fully nested within the meta-model's outer fold splits, so the conclusion of "no headroom" should be viewed as conservative. We examined this bias with single-nested-seed re-runs, but fully nested base feature generation across multiple random seeds remains future work. The results come from a consolidated 5-fold OOF experiment; bootstrapping over authors rather than folds, random seeds, or data splits means the confidence intervals do not account for base-model training variance across different data splits, although sensitivity to split design was assessed via 20 random-seed re-runs. The resampling scale ($n_{\text{null}} = 200$, $n_{\text{boot}} = 2000$) supports the one-sided exclusion test for "no positive headroom"; however, with paired-resampling standard errors incorporating bootstrap–null covariance, only ZS+BEATs establishes TOST equivalence at $\delta = 0.01$. Thus, the other scenarios report limited headroom rather than proven equivalence. The cross-validation (CV) versus leaderboard comparison uses only $n = 3$ author-replicated single submissions and diagnoses performance gaps or rank inversions rather than rank correlation.

The evaluation scope is also limited. The soundscape evaluation is small-scale (66 audio files, 1,478 time windows, and 43–47 shared classes), and the robustness-gap analysis used only the "fold-4-held-out" checkpoint for each model family; exhaustive multi-seed testing remains future work. The pseudo-labeling null test is power-limited because it used only a weak teacher model (the Top-1 result from Perch-ZS) and a single random seed. Thus, the result means only that no improvement headroom was detected under this specific "clean" setup; it does not imply that pseudo-labeling cannot help. Stronger teacher models, such as CNN ensembles or calibrated soft-fusion stacked models, across multiple random seeds are a logical next step. BEATs was the weakest-performing model family (macro-AUC 0.887), so its "linear-probe-anchor" negative headroom may partly reflect model limitations; however, stronger challengers – CNNs and PANNs – also failed to show positive headroom, so the overall "null" was not solely attributable to BEATs. On the soundscape target task, the frozen focal-learned combiner failed to outperform the Perch V2 zero-shot anchor, so no positive deployment gain can be claimed; a comprehensive soundscape transfer experiment over all eight cells remains future work.

The construction of metrics, priors, and masks is also critical. Macro-AUC was calculated and verified

Table 5
Run and submission manifest. The 2026-06-08 submissions are post-deadline (late) reruns and did NOT change the official competition rank (262/4085); validation-track configurations were pre-registered and each submitted once. CV is focal recordist-grouped 5-fold and is not comparable to soundscape LB across rows. **Bold** = best pub/priv LB within track.

Date	Sub-id	Config	Owner	Track	CV	Pub LB	Priv LB
2026-05-31	53225430	Bronze stack	C	competition	–	**0.95044**	0.94221
2026-05-29	53155487	EoS8 best	C	competition	–	0.95004	**0.94247**
2026-06-08	53467777	ZS+CNN fusion	A	validation	0.949	**0.85903**	**0.87930**
2026-06-08	53467908	Perch ZS	A	validation	0.936	0.82380	0.84193
2026-06-08	53468008	CNN fold-4	A	validation	0.940	0.79782	0.80973
2026-06-08	53466774	Late rerun	C	late test	–	0.95045	0.94224

from the implementation code rather than copied from real-time Kaggle dashboard data; the duplicate-anchor control difference was set to zero by construction. The ecological spatial prior used global iNaturalist data rather than Pantanal test-site coordinates, so its negative effect reflects geographical mismatch and does not invalidate "location-matched priors." Parameters α and β were preset and not hyperparameter-tuned. The common "evaluable-class mask" excluded zero-positive and long-tail classes, so potential-improvement estimates apply to the mask ($\approx$ 194–197 classes) rather than all 234 classes (see the "Methods" section). Prevalence stratification (Table 4) indicated that masking did not make "no improvement" structurally inevitable, and excluded long-tail classes were reviewed separately. Overall, the study covers a single task, a single anchor, and a single OOF (out-of-fold) prediction matrix; conclusions about potential improvement are specific to the Perch V2 anchor on the Pantanal dataset, the single BirdCLEF++ task, and the single aggregated OOF result. Because the de-biased intervals use a conservative paired-resampling estimator, they should be viewed as conservative bound estimates rather than precise confidence intervals. We have not verified whether "no detectable improvement" or the "coverage-wall" applies to other bioacoustic foundation anchors (e.g. BirdNET, SurfPerch, or AVES); replicating the audit with a second foundation model anchor is future work.

7. Data, Ethics, and Reproducibility

The data are the BirdCLEF++ 2026 focal Xeno-Canto training recordings and Pantanal soundscape labels, used under Kaggle competition rules; this working note is released under the CC BY 4.0 license. The task involves passive acoustic monitoring and no human subjects. Community-sourced components are attributed and comply with their licenses; pseudo-labels are treated as model outputs rather than ground truth. Each cross-validation (CV) number is linked to artifacts, execution commands, and random seeds. The reproducible anchor is the Perch-V2 CPU-ONNX head with trained CNN, PANNs, and BEATs sidecars. The audit harness is released publicly at https://github.com/HaochenSHI66/birdclef2026-audit, containing fixed folds, cached out-of-fold predictions, per-cell result JSON files, scoring, bootstrapping, label-permutation, and TOST code, plus a runnable minimal baseline. Community components (Perch V2, PANNs, BEATs, and the competition stack) are cited and attributed rather than redistributed.

8. Conclusion

For BirdCLEF++, we audited whether supervised challengers and ensemble heuristics add measurable accuracy once a robust bioacoustic foundation anchor is already present, using "clean reruns," ablations, and provenance checks rather than model-novelty claims. Methodologically, we combined main effects with quantitative uncertainty—author-clustered bootstrapped confidence intervals, permutation-test nulls, and one-sided exclusion tests—and released the audit framework. The coverage audit found that 31 of 234 scored classes fell outside Perch's fixed eBird output label space, a limitation concentrated

almost entirely in Insecta, while vocabulary-covered classes with positive samples generally achieved high AUC transfer performance [3, 6, 7]. The oracle-headroom audit found no detectable in-domain hard-selection gain over the best single model across eight cells, including the rare/difficult stratum: a power-bounded rejection of practically sized hard-selection gains [9], not a claim that the models are interchangeable. Calibrated soft fusion was the only effective combined signal for focal-CV data. Post-processing and pseudo-labeling showed no detectable improvement in this audit, while focal-to-soundscape shift was the primary error source, with Perch V2 embeddings showing the smallest degradation in this diagnostic.

The practical implication is to shift R&D away from focal-CV hard routing and combinatorial heuristics, and toward the factors driving error. For the unmapped taxa, that means output paths beyond Perch's fixed eBird head, supported by local sonotype labels, target-domain pseudo-labels, or external recordings. For model selection, it means validating on soundscape-like data before trusting focal-CV rank or a learned combiner. Future work should use base features from multiple random seeds and fully nested, outer-fold aligned configurations; test multi-seed pseudo-labeling with stronger teachers; repeat the audit with a second foundation-model anchor; incorporate genuine location-specific ecological priors; and, most importantly, improve focal-to-soundscape domain adaptation.

Declaration on Generative AI

The research questions, the reproducible-audit framing, the experimental designs, and the integrity and verification standards applied throughout this work were conceived, specified, and directed by the author. The author used Claude, ChatGPT, and GitHub Copilot, under the author's direction, for grammar and style polishing, LaTeX and table editing, and code-audit assistance. No AI tool generated experimental data, results, or claims; all numbers come from the author's runs and all scientific claims are the author's. Every reported number was independently checked by the author against its source artifact (logs, result files, and recorded commands/seeds), and all scientific claims, interpretations, and conclusions are the author's own. The author takes full responsibility for the accuracy, integrity, and reproducibility of the work.

References

[1] L. Picek, L. Adam, S. Kahl, R. Bossy, L. Chrobak, H. Goëau, K. Papafitsoros, H. Klinck, W.-P. Vellinga, R. Planqué, T. Denton, K. Barnard, C. Nédellec, L. Deléger, M. Courtin, G. Martellucci, I. Moummad, F. Vinatier, P. Bonnet, A. Joly, Overview of LifeCLEF 2026: AI challenges for biodiversity understanding and ecosystem management, in: International Conference of the Cross-Language Evaluation Forum for European Languages (CLEF), Springer, 2026.

[2] S. Kahl, T. Denton, L. Sugai, L. Piatti, W. H. Arruda Moreno, M. M. Barbosa, M. Eibl, C. Z. Fieker, C. M. Garcia, D. L. Hokama Sousa, J. E. d. A. Júnior, R. C. Kridler, M. Lasseck, A. V. d. Melo, H. Müller, M. d. O. Neves, M. G. d. Reis, L. K. Sampaio Teles, K.-L. Schuchmann, J. L. M. M. Sugai, K. T. P. Azevedo, P. d. N. Lopes, M. I. Marques, H. Klinck, H. Glotin, H. Goëau, W.-P. Vellinga, R. Planqué, A. Joly, Overview of BirdCLEF+ 2026: Acoustic species identification in the Pantanal, South America, in: Working Notes of CLEF 2026 – Conference and Labs of the Evaluation Forum, 2026.

[3] J. S. Cañas, S. Kahl, T. Denton, M. P. Toro-Gómez, S. Rodriguez-Buritica, J. L. Benavides-Lopez, J. S. Ulloa, P. Caycedo-Rosales, H. Klinck, H. Goëau, W.-P. Vellinga, R. Planqué, A. Joly, Overview of BirdCLEF+ 2025: Multi-taxonomic sound identification in the middle magdalena, colombia, in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, volume 4038 of *CEUR Workshop Proceedings*, 2025. URL: https://ceur-ws.org/Vol-4038/paper_232.pdf, full author list verified against CEUR Vol-4038 paper_232.

[4] A. Miyaguchi, M. Gustineli, A. Cheung, Distilling spectrograms into tokens: Fast and lightweight bioacoustic classification for BirdCLEF+ 2025, in: Working Notes of CLEF 2025 – Conference and

Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025. URL: https://arxiv.org/abs/2507.08236, dS@GT; arXiv:2507.08236.

[5] B. van Merriënboer, V. Dumoulin, J. Hamer, L. Harrell, A. Burns, T. Denton, Perch 2.0: The bittern lesson for bioacoustics, arXiv preprint arXiv:2508.04665 (2025). URL: https://arxiv.org/abs/2508.04665.

[6] D. Robinson, M. Miron, M. Hagiwara, B. Weck, S. Keen, M. Alizadeh, G. Narula, M. Geist, O. Pietquin, NatureLM-audio: An audio–language foundation model for bioacoustics, in: Proceedings of the International Conference on Learning Representations (ICLR), 2025. URL: https://arxiv.org/abs/2411.07186, arXiv:2411.07186.

[7] Y. Chen, N. Lang, B. C. Schmidt, A. Jain, Y. Basset, S. Beery, M. Larrivée, D. Rolnick, Open-insect: Benchmarking open-set recognition of novel species in biodiversity monitoring, arXiv preprint arXiv:2503.01691 (2025). URL: https://arxiv.org/abs/2503.01691.

[8] M. Ojala, G. C. Garriga, Permutation tests for studying classifier performance, Journal of Machine Learning Research 11 (2010) 1833–1863. URL: https://www.jmlr.org/papers/volume11/ojala10a/ojala10a.pdf.

[9] G. C. Cawley, N. L. C. Talbot, On over-fitting in model selection and subsequent selection bias in performance evaluation, Journal of Machine Learning Research 11 (2010) 2079–2107. URL: https://www.jmlr.org/papers/volume11/cawley10a/cawley10a.pdf.

[10] L. Rauch, R. Schwinger, M. Wirth, R. Heinrich, D. Huseljic, M. Herde, J. Lange, S. Kahl, B. Sick, S. Tomforde, C. Scholz, BirdSet: A large-scale dataset for audio classification in avian bioacoustics, in: Proceedings of the International Conference on Learning Representations (ICLR), 2025. URL: https://arxiv.org/abs/2403.10380, arXiv:2403.10380.

[11] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, F. Wei, BEATs: Audio pre-training with acoustic tokenizers, in: Proceedings of the 40th International Conference on Machine Learning (ICML), 2023. URL: https://arxiv.org/abs/2212.09058, arXiv:2212.09058.

[12] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, M. D. Plumbley, PANNs: Large-scale pretrained audio neural networks for audio pattern recognition, IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020) 2880–2894. doi:10.1109/TASLP.2020.3030497, arXiv:1912.10211.

[13] B. Williams, B. van Merriënboer, V. Dumoulin, J. Hamer, E. Triantafillou, A. B. Fleishman, M. McKown, J. Munger, A. N. Rice, A. Lillis, C. White, C. Hobbs, T. Razak, D. Curnick, K. E. Jones, T. Denton, Leveraging tropical reef, bird and unrelated sounds for superior transfer learning in marine bioacoustics, Philosophical Transactions of the Royal Society B 380 (2025) 20240280. doi:10.1098/rstb.2024.0280, surfPerch; arXiv:2404.16436.

[14] M. Hagiwara, AVES: Animal vocalization encoder based on self-supervision, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023. URL: https://arxiv.org/abs/2210.14493, arXiv:2210.14493.

[15] Y. Gong, Y.-A. Chung, J. Glass, AST: Audio spectrogram transformer, in: Proc. Interspeech 2021, 2021, pp. 571–575. doi:10.21437/Interspeech.2021-698, arXiv:2104.01778.

[16] R. Schwinger, P. Vali Zadeh, L. Rauch, M. Kurz, T. Hauschild, S. Lapp, S. Tomforde, Foundation models for bioacoustics – a comparative review, Ecological Informatics 96 (2026) 103765. doi:10.1016/j.ecoinf.2026.103765, arXiv:2508.01277.

[17] M. Hagiwara, B. Hoffman, J.-Y. Liu, M. Cusimano, F. Effenberger, K. Zacarian, BEANS: The benchmark of animal sounds, arXiv preprint arXiv:2210.12300 (2022). URL: https://arxiv.org/abs/2210.12300.

[18] R. Heinrich, L. Rauch, B. Sick, C. Scholz, AudioProtoPNet: An interpretable deep learning model for bird sound classification, Ecological Informatics 87 (2025) 103081. doi:10.1016/j.ecoinf.2025.103081, arXiv:2404.10420.

[19] S. Kahl, T. Denton, H. Klinck, H. Glotin, H. Goëau, W.-P. Vellinga, R. Planqué, A. Joly, et al., Overview of BirdCLEF 2023: Automated bird species identification in eastern africa, in: Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum, volume 3497 of *CEUR Workshop Proceedings*, 2023. URL: https://ceur-ws.org/Vol-3497/paper-164.pdf.

[20] S. Kahl, T. Denton, H. Klinck, V. Ramesh, V. Joshi, H. Glotin, H. Goëau, W.-P. Vellinga, R. Planqué,

A. Joly, et al., Overview of BirdCLEF 2024: Acoustic identification of under-studied bird species in the western ghats, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2024. URL: https://hal.science/hal-04719578.

[21] A. Miyaguchi, N. Zhong, M. Gustineli, C. Hayduk, Transfer learning with semi-supervised dataset annotation for birdcall classification, in: Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum, volume 3497 of *CEUR Workshop Proceedings*, 2023. URL: https://arxiv.org/abs/2306.16760, dS@GT BirdCLEF 2023; arXiv:2306.16760.

[22] A. Miyaguchi, A. Cheung, M. Gustineli, A. Kim, Transfer learning with pseudo multi-label birdcall classification for ds@gt BirdCLEF 2024, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2024. URL: https://arxiv.org/abs/2407.06291, dS@GT BirdCLEF 2024; arXiv:2407.06291.

[23] V. Sydorskyi, F. Gonçalves, Tackling domain shift in bird audio classification via transfer learning and semi-supervised distillation: A case study on BirdCLEF+ 2025, in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, volume 4038 of *CEUR Workshop Proceedings*, 2025. URL: https://ceur-ws.org/Vol-4038/paper_256.pdf, birdCLEF+ 2025 2nd place; authors confirmed via project repo BibTeX.

[24] L. I. Kuncheva, C. J. Whitaker, Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy, Machine Learning 51 (2003) 181–207. doi:10.1023/A:1022859003006.

[25] D. H. Wolpert, Stacked generalization, Neural Networks 5 (1992) 241–259. doi:10.1016/S0893-6080(05)80023-1.

[26] A. Niculescu-Mizil, R. Caruana, Predicting good probabilities with supervised learning, in: Proceedings of the 22nd International Conference on Machine Learning (ICML), 2005, pp. 625–632. doi:10.1145/1102351.1102430.

[27] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks, in: Proceedings of the 34th International Conference on Machine Learning (ICML), 2017, pp. 1321–1330. ArXiv:1706.04599.

[28] R. Schwinger, B. McEwen, V. S. Kather, R. Heinrich, L. Rauch, S. Tomforde, Uncertainty calibration of multi-label bird sound classifiers, arXiv preprint arXiv:2511.08261 (2025). URL: https://arxiv.org/abs/2511.08261.

[29] D. J. Schuirmann, A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability, Journal of Pharmacokinetics and Biopharmaceutics 15 (1987) 657–680. doi:10.1007/BF01068419.

[30] D. Lakens, Equivalence tests: A practical primer for t tests, correlations, and meta-analyses, Social Psychological and Personality Science 8 (2017) 355–362. doi:10.1177/1948550617697177.

[31] D. Lakens, A. M. Scheel, P. M. Isager, Equivalence testing for psychological research: A tutorial, Advances in Methods and Practices in Psychological Science 1 (2018) 259–269. doi:10.1177/2515245918770963.

[32] R. Heinrich, L. Rauch, B. Sick, C. Scholz, Adversarial training improves generalization under distribution shifts in bioacoustics, arXiv preprint arXiv:2507.13727 (2025). URL: https://arxiv.org/abs/2507.13727.

[33] M. Lasseck, Improving bird recognition using pseudo-labeled recordings from the target location, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, volume 3740 of *CEUR Workshop Proceedings*, 2024, pp. 2110–2118. URL: https://ceur-ws.org/Vol-3740/paper-199.pdf.

[34] E. Witting, H. de Heer, J. Lim, C. T. Kopar, K. Sándor, Addressing the challenges of domain shift in bird call classification for BirdCLEF 2024, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, volume 3740 of *CEUR Workshop Proceedings*, 2024, pp. 2244–2259. URL: https://ceur-ws.org/Vol-3740/paper-211.pdf.

[35] L. Hong, Domain adaption for birdcall recognition: Progressive knowledge distillation with semi-supervised and self-supervised soundscape labeling, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, volume 3740 of *CEUR Workshop Proceedings*, 2024, pp. 2103–2109. URL: https://ceur-ws.org/Vol-3740/paper-198.pdf.

[36] K. Dmitriev, Methods for training convolutional neural networks to identify bird species in complex soundscape recordings, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, volume 3740 of *CEUR Workshop Proceedings*, 2024, pp. 2044–2052. URL: https://ceur-ws.org/Vol-3740/paper-193.pdf.

[37] J. Tan, A. Wang, Dual-branch network for species identification via passive acoustic monitoring, in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, volume 4038 of *CEUR Workshop Proceedings*, 2025, pp. 3248–3253. URL: https://ceur-ws.org/Vol-4038/paper_259.pdf.

[38] S. S. Gokulnath, C. Das, A. Gaikwad, K. Senthilnathan, S. P. Sawant, One detector per bird: A scalable binary classification approach for BirdCLEF+ 2025, in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, volume 4038 of *CEUR Workshop Proceedings*, 2025, pp. 3179–3198. URL: https://ceur-ws.org/Vol-4038/paper_254.pdf.

[39] A. Sampath Kumar, T. Schlosser, D. Kowerko, TUC media computing at BirdCLEF 2024: Improving birdsong classification through single learning models, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, volume 3740 of *CEUR Workshop Proceedings*, 2024, pp. 2201–2208. URL: https://ceur-ws.org/Vol-3740/paper-206.pdf.

[40] Y. McLatchie, A. Vehtari, Efficient estimation and correction of selection-induced bias with order statistics, Statistics and Computing 34 (2024) 132. doi:10.1007/s11222-024-10442-4, arXiv:2309.03742.

[41] S. Holm, A simple sequentially rejective multiple test procedure, Scandinavian Journal of Statistics 6 (1979) 65–70. URL: https://www.jstor.org/stable/4615733.

[42] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, S. Belongie, The iNaturalist species classification and detection dataset, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8769–8778. URL: https://arxiv.org/abs/1707.06642, arXiv:1707.06642.

[43] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, A. Smola, A kernel two-sample test, Journal of Machine Learning Research 13 (2012) 723–773. URL: https://www.jmlr.org/papers/volume13/gretton12a/gretton12a.pdf.