Mapping the Public Ceiling: Negative Results and a Platform-Constraint Hypothesis for BirdCLEF+ 2026

Wei Meng Soh[1]

[1] *Whyme Labs (independent), Malaysia*

Abstract

Most competition working notes describe a winning recipe. This note deliberately does the opposite, and argues the inversion is useful: a systematic, measurement-driven account of *why a strong shared public baseline could not be beaten with consumer-scale resources*, and what that reveals about the structure of modern bioacoustic competitions. By mid-competition a single architectural pattern — a frozen Perch v2 embedding model, a Light-ProtoSSM sequence head, a distilled Sound-Event-Detection (SED) CNN, and a taxonomy-smoothing post-processor — had been openly shared and converged on a public-leaderboard score of **0.950**, where (as of the public-LB snapshot on 2 June 2026) roughly **890 of 4243 teams scored in the** $[0.950, 0.951)$ **band**. We treat this plateau as an object of study rather than a target. We used a disciplined *local cross-validation gate* — a leak-controlled offline split scored before spending one of five daily submission slots — to test every lever available to a small team: own-trained SEDs, foundation-model linear probes, rare-taxon specialists, post-hoc calibration tricks, and architecturally-diverse model ensembles. Each was neutral or negative. Two observations we found practically consequential in this setting: (1) the public components' apparent near-perfect held-out accuracy is *optimistic by construction* — they are fit on the only in-domain labelled audio available for validation — which inverts the most natural offline comparison and makes held-out macro-AUC an unreliable basis for ranking a challenger against the baseline; and (2) the binding constraint for a small team is not modeling skill but **platform limits** — a $\sim$1 MB notebook-size cap and a 90-minute CPU runtime budget — which physically forbid the one technique with genuine headroom (ensembling multiple full foundation-model pipelines). On a single held-out split, architecturally-diverse CNN ensembling improved macro-AUC by **+0.021** over its best member (a leak-controlled, directional result, §5.6), yet that ensemble could not be deployed within the runtime envelope. We close with a hypothesis, offered explicitly as conjecture since the leaders' solutions are private: in the foundation-model era a public pipeline assembled from large pretrained components plus a community-discovered post-processing trick is a remarkably hard floor, and we conjecture the residual gap to the leaders is paid largely in up-front training compute rather than inference-time cleverness. All code and the offline-gate harness are released openly.

Keywords

bioacoustics, BirdCLEF, foundation models, negative results, ensemble diversity, knowledge distillation, data leakage

1. Introduction

BirdCLEF+ 2026, the bird-focused task of the LifeCLEF 2026 evaluation campaign [1, 2], asked participants to identify which of 234 species — birds, amphibians, mammals, reptiles, and insects — vocalize in 1-minute soundscapes from the Brazilian Pantanal, an important conservation-monitoring task. Submissions are CPU-only Kaggle notebooks evaluated on a hidden set of $\sim$600 recordings, scored by macro-averaged ROC-AUC that skips classes with no true positives in the test set; the runtime budget is **90 minutes**, internet is disabled, and each team may submit five times per day.

The competition exhibited an unusually sharp *public-kernel monoculture*. A lineage of openly-shared notebooks ("Ensemble-of-Solutions", EoS.x, and their forks) raised the public ceiling in discrete, community-wide jumps — $0.928 \to 0.948 \to 0.949 \to$ **0.950** — each jump driven by one new mechanism that the field adopted within days. The cumulative public-LB distribution (snapshot 2 June 2026, 4243 teams) was top-heavy and then sharply flat: **11 teams** ≥ 0.96, **42** ≥ 0.955, **252** ≥ 0.951, and **1140** ≥ 0.950 — i.e. $\sim$890 teams whose best public score fell in the narrow $[0.950, 0.951)$ band, the

CLEF 2026: Conference and Labs of the Evaluation Forum, September 21–24, 2026, Jena, Germany

wmhy.tech@gmail.com (W. M. Soh)

large population running the shared pipeline essentially unchanged. (Public-LB counts drift as teams resubmit; we report one dated snapshot.)

We entered this monoculture late and spent the competition on one question: *can a small team, on a single consumer GPU plus a small cloud budget, add anything to the public 0.950 pipeline that survives contact with the hidden test?* Our answer is negative. Each hypothesis was reduced to a measured quantity and gated before a submission was spent. The contribution of this note is therefore not a model but a **map**: what we tried, the numbers, the structural reason each path failed, and a small reusable methodology for resource-constrained competitors. Negative results reported with measurements are comparatively rare in this venue, since most write-ups come from teams that succeeded; we hope the surrounding terrain is useful.

2. Related work

BirdCLEF has run annually since 2014; recent editions converged on transfer-learning from large audio/bird foundation models followed by sequence modeling and heavy ensembling of diverse CNNs (the consistent ingredient in winning solutions across 2023–2025) [3]. Our public baseline reflects this: Perch v2 [4] as the embedding backbone, a ProtoSSM-style selective-state-space [5] head over its embeddings, and a distilled SED CNN. Our own-model attempts draw on noisy-student self-training [6] (pseudo-labeling unlabeled soundscapes with a teacher, then training a student) and on foundation-model linear probing (Bird-MAE [7], AudioMAE [8]). Two themes we engage critically are (i) knowledge distillation's ceiling [9] — a student trained on a teacher's outputs is bounded by that teacher — and (ii) ensemble diversity — that decorrelated architectures, averaged, beat any single member. Our contribution relative to this literature is empirical and negative: we quantify *where each of these reliably-useful techniques stops working* under this competition's specific data-leakage and platform constraints.

3. Anatomy of the public 0.950 pipeline

The converged public solution is a four-stage rank ensemble:

1. **Perch v2** [4] — Google's bird-vocalization foundation model — produces a 1536-d embedding and 234-logit score per 5-second window. It is the single most expensive inference stage and the dominant signal.
2. **Light-ProtoSSM** — a small (≈ 0.75–5.8 M-parameter) bidirectional selective-state-space [5] head with prototype classification and cross-window attention — is fit (or loaded pre-trained) on Perch embeddings of the labelled soundscapes, contributing $\sim 60\%$ of the rank blend.
3. **Distilled SED** — a 5-fold EfficientNet-B0 sound-event detector on 256×313 log-mel input, distilled [9] from a stronger teacher ensemble, contributing $\sim 40\%$.
4. **`TAX_SMOOTHING`** — the final-jump mechanism ($0.949 \rightarrow 0.950$): each species' score is pulled slightly toward the mean of its genus- and class-siblings ($\alpha_{\text{genus}} \approx 0.15$, $\alpha_{\text{class}} \approx 0.05$), exploiting taxonomic correlation among co-occurring Pantanal taxa.

We verified directly that every public 0.950 notebook is a scalar-tuned variant of this stack: forks differing in post-processing (BirdNET sidecars, PCEN, out-of-fold gating, iterative smoothing) and even one with a different ProtoSSM implementation all returned exactly 0.950 or below.

4. Method: local cross-validation as a pre-submission gate

The scarce resources here are not FLOPs but **submission slots** (five per day, each scored only after a multi-hour hidden-test rerun) and **trustworthy validation signal** (the labelled audio is tiny and, as §5.2 shows, optimistically biased toward the public models). Our central methodological move was to

never spend a slot on an unmeasured hypothesis: for each idea we built a cheap offline test on cloud GPU (Modal) or the local GPU and used it as a go/no-go *gate* on the submission budget.

4.1. The offline split, and how it is kept leak-controlled

At bottom this is a *local cross-validation split* of the kind familiar to Kaggle competitors — hold out labelled data, score a candidate on it offline, and promote it to a real submission only if it clears a bar. What we add is (i) treating that split as a *hard gate* on a five-per-day budget rather than a diagnostic weighed against intuition, and (ii) making it *leak-aware*, because in this competition the most natural local split silently favours the public baseline (§5.2).

Concretely, the split is **file-level, not window-level**: we held out 13 whole labelled soundscapes and used *no* 5-second window from any held-out recording anywhere in training. For our own models this exclusion is enforced at *every* stage, including the soundscape pseudo-labelling that generates noisy-student targets — so neither a student nor the teacher that labels its training data ever sees a held-out file. This is what makes the ensemble-vs-member comparison of §5.6 clean: both sides of that comparison are equally held out from the same 13 files. The one exclusion we *cannot* enforce is on the public components: they were trained by others on the full labelled set, so when we score them on our hold-out their numbers are leak-optimistic by construction. That asymmetry — our side genuinely held out, the baseline's side not — is the validation trap of §5.2, and it is why we lean on *relative*, leak-controlled gates (ensemble-vs-best-single; per-class win counts against the public SED) rather than absolute comparisons to the baseline.

For each idea we instantiated one of three gates:

- **`compare_to_teacher`** — train a candidate SED with the clean *file-level* hold-out of 13 labelled soundscapes; run the public distilled SED on the *same* held-out files; report both macro-AUCs and the per-class win count.
- **`eval_*_probe`** / **`eval_rare_taxa`** — train linear probes on frozen foundation-model embeddings; evaluate against the public SED *sliced to specific taxa*; measure orthogonality (how many classes the candidate strictly beats the baseline on).
- **`eval_sed_ensemble`** — measure whether an ensemble's macro-AUC exceeds its best single member (a leak-free decorrelation test, independent of any public teacher).

Each gate ran in minutes on a single cloud GPU at low cost (we did not log a per-gate cost breakdown; total cloud spend across the study was a few US dollars). Each falsified direction would otherwise have consumed one-to-many submission slots and a day of latency. We regard this leak-aware gating discipline (§9) as a small transferable contribution for code-competitions with daily quotas and slow scoring.

4.2. An agentic experimentation workflow

This study was conducted as a human-directed, AI-agent-executed workflow: a coding agent (Anthropic's Claude, driven via Claude Code) carried out the loop of writing each gate, launching cloud (Modal) training and evaluation jobs, parsing the results, and proposing the next hypothesis, under human review at each decision point. The discipline the agent enforced is exactly what made a five-slots-per-day budget tractable: every hypothesis was turned into a sub-dollar offline measurement and *gated* before a submission was spent, and dead ends were recorded with numbers rather than re-litigated by intuition. The agent also caught failures a human might have shrugged off — for example, it surfaced that our compare-to-teacher comparison was leak-biased (§5.2) only because it logged and diffed the per-class win counts. We document this because the organizers explicitly invited participants to describe AI-assisted and autonomous workflows; in our case the agentic loop was not a coding convenience but the mechanism that let a small team exhaustively map the ceiling within the platform's constraints. All experiments were designed under human direction and every reported number was manually verified (see the AI-use disclosure before the references).

Table 1
Levers tested against the public 0.950 baseline. ("LB" = public leaderboard at submission time; held-out figures are single-split offline gates, §4.)

Lever	Measured outcome	Verdict	Mechanism of failure
Own SED, focal-trained	LB 0.931	-0.019	distilled copy of a weaker teacher
Own SED, soundscape noisy-student	LB 0.933	-0.017	bounded by the public SED's own pseudo-labels
`compare_to_teacher` (clean 13-file hold-out)	student 0.910 vs public **0.994**	—	but the 0.994 is leak-optimistic (§5.2)
Bird-MAE / AudioMAE linear probe	beats public on **0/27** classes	dominated	no orthogonal signal to blend
Rare-taxon (frog/insect) specialist	beats public on **0/16** rare classes	dominated	public SED already 0.991 on rare taxa
Post-hoc tricks (BirdNET, co-occurrence, genus-proxy, iterative smooth)	LB 0.948–0.950	neutral/neg.	perturb a tight calibrated equilibrium
Diverse-CNN SED ensemble (4 archs)	**0.954** vs best single **0.933** (single split)	**+0.021**	gain real but undeployable (§6)
Diverse-ensemble *swapped* into pipeline	LB 0.945	-0.005	our members individually $<$ public SED on hidden test
Diverse-ensemble *added* (7 or 9 SEDs)	runtime timeout	—	90-min CPU budget exceeded

5. Systematic negative results

Table 1 consolidates the experiments. Each row was decided by a gate (offline macro-AUC, leak-aware) and/or a leaderboard submission.

5.1. Own-trained SED cannot match the distilled public SED

We trained EfficientNet-B0 SEDs two ways: focal-recording-only (public LB **0.931**) and a noisy-student [6] variant on 10,592 pseudo-labelled soundscapes $\rightarrow$ 127,896 soft-labelled windows (public LB **0.933**). Both regressed $\sim$0.017–0.019 below the baseline when blended in. The gate indicated the cause: on the 13-file hold-out the public SED scored **0.994** versus our student's **0.910** — though that 0.994 is optimistically biased (§5.2). A single own-trained CNN on one GPU does not reach a model distilled from a strong teacher ensemble — and, because our students learn from the public SED's *own* pseudo-labels, they are its distilled copies, structurally bounded below it [9].

5.2. The validation trap: optimistic held-out for public components

The gate number above (public SED 0.994 on hold-out) is **misleading in a way worth making explicit**, because it determined how we validated everything afterward. The public Light-ProtoSSM is, by inspection of the public notebook, fit on the labelled soundscapes (it trains on them in-kernel); the distilled SED was plausibly distilled with them in scope. Either way, the public pipeline's 0.994 on a held-out *subset of those same soundscapes* is optimistic by construction, whereas our genuinely-held-out student is scored fairly — an asymmetric comparison. The gap between the public components' $\sim$0.99 on labelled audio and the full pipeline's 0.950 / 0.942 on the hidden test is consistent with this optimism. The practical consequence: **an offline comparison of a challenger against a public component on the labelled soundscapes is biased in the baseline's favour**, so held-out macro-AUC cannot reliably rank a small team's model against the shared pipeline. This is, of course, an instance of the textbook train/test-leakage pitfall; our contribution is not the concept but the demonstration that, in

this competition, it silently invalidates the *most natural* validation a competitor would reach for, and the consequent discipline of preferring *leak-controlled relative* gates (e.g. ensemble-vs-best-single, §5.6) over absolute comparisons to the baseline. (Whether the precise mechanism is label leakage or in-domain over-fit plus distribution shift, the methodological conclusion is identical.)

5.3. Foundation-model probes are strictly dominated

Linear probes on Bird-MAE-Base [7] and AudioMAE [8] embeddings beat the public SED on **0/27** held-out classes with positives; every rank-blend weight monotonically *reduced* macro-AUC. A strictly-dominated source cannot improve a calibrated ensemble — it can only dilute it.

5.4. Rare-taxon specialists: a plausible premise, falsified

We hypothesized the bird-tuned pipeline would be weak on rare *Amphibia/Insecta* — stationary frog/insect textures unlike bird syllables — and that a texture-oriented model (AudioMAE [8]) could win there. Sliced to 16 measurable rare classes (954 held-out positives), the gate refuted it: the public SED scored **0.991** on exactly these taxa (it had learned them from the labelled soundscapes), and our specialists beat it on **0/16** (Bird-MAE 0.876, AudioMAE 0.799 macro-AUC on the slice). External audio cannot rescue the data-starved tail either: of 53 classes with < 10 training recordings, Xeno-Canto (a bird archive) covered 1, and 28 — including the `sonXX` insect "morphotypes" — are competition-internal labels with no scientific name and thus no external data anywhere.

5.5. Post-hoc tricks perturb a tight equilibrium

Genus-proxy boosting, sonotype mirroring, per-taxon temperature, BirdNET 3-way blending, cross-class co-occurrence boosting, and iterative second-pass smoothing each layered onto the 0.950 base returned 0.948–0.950 (neutral-to-negative). The public pipeline's calibration chain (priors, rank-aware scaling, power optimization, threshold tuning, taxonomy smoothing) is a tight equilibrium; post-hoc reorderings disturb it more than they help. Tellingly, the community's own last $+0.001$ came not from *adding* a trick but from `TAX_SMOOTHING` applied *inside* the calibrated chain at the right scale.

5.6. Diverse-CNN ensembling: a real gain that cannot be deployed

The most reliable technique in the field is ensembling architecturally-diverse CNNs, and we used it as a *controlled internal* test rather than a comparison to the (leak-flattered) baseline. We trained four diverse SED backbones — EfficientNetV2-S, ConvNeXt-Small, EfficientNet-B3, SEResNeXt50. Crucially, the 13 hold-out files were excluded from **all** four members' training, *including the soundscape pseudo-label generation* (§4.1) — so the ensemble-vs-member comparison is leak-controlled with respect to that split (both sides are equally held out, unlike the baseline of §5.2). On that single split the four-model rank ensemble scored **0.9542** versus the best single member's **0.9332**, a **+0.021** improvement — the expected diversity effect, observed cleanly. We stress this is a *single-split point estimate*: we did not run multiple folds or seeds and report no confidence interval, so $+0.021$ should be read as directional corroboration of the well-known diversity benefit, not a precise effect size (see §7).

This still could not break 0.950, for two largely independent reasons that are the crux of this note. First, *deployment runtime*: adding the four SEDs to the public pipeline ($5 + 4 = 9$, or even $5 + 2 = 7$ inferences) exceeded the 90-minute budget and timed out — a wall independent of model quality. Second, *substitution cost*: swapping two public SEDs for two of ours (holding the count at 5, runtime-safe) scored **0.945** on the LB. We read this single point cautiously: it is *consistent with* our members being individually weaker than the public SED on the genuine hidden test (where the baseline's leak advantage vanishes), but it is confounded — the rank-blend weights were not re-tuned for the substitution, and our members were only half-trained when a cloud-budget cap froze them, which independently predicts a drop. The runtime conclusion is therefore firm; the substitution conclusion is suggestive, not proven. What the two together establish is narrower but solid: *we could not realize the diversity gain within the*

platform envelope, whether because of runtime (certain) or insufficient member quality (likely, given half-training) or both.

6. The platform-constraint finding

The deeper lesson generalizes beyond SEDs. The one strategy with real headroom over a converged baseline is to **ensemble multiple full, diverse pipelines**. We attempted to combine two genuinely different public 0.950 pipelines (distinct ProtoSSM implementations and auxiliary models) in one notebook. This is blocked by *platform physics*, not modeling:

- **Notebook-size cap ($\sim$1 MB).** The two full pipelines concatenated to 1.08 MB and were rejected at upload; a single pipeline is 0.77 MB.
- **Runtime cap (90 min CPU).** Each pipeline's Perch pass nearly fills the budget; two passes time out — the same wall that killed our additive-SED ensembles.

These limits are, we argue, *a* structural reason for the monoculture: heavy foundation-model pipelines cannot be ensembled at inference inside the platform, so the public field — which mostly forks the same heavy pipeline — is funnelled into a single 0.950 cluster. We then *conjecture* (the leaders' solutions are private, so this is not established) that the teams above 0.95 escape the funnel by pre-training their own lighter models that are individually competitive yet cheap enough to run several within 90 minutes, paying the cost up front in training compute that a late, small entrant on one consumer GPU cannot match. This conjecture has at least an empirical prior: published BirdCLEF 2023–2025 winning solutions [3] consistently rely on *ensembles of several independently-trained CNNs* run within the inference budget, rather than a single shared backbone. We present it as the most parsimonious hypothesis consistent with our measurements, not as a demonstrated fact about the 11 teams ≥ 0.96.

7. Threats to validity

We state the limits of these conclusions plainly. (i) **Single-split, no error bars.** Our offline gates use one fold and a small hold-out (13 files; 32 classes with positives for the absolute gates, 16 rare classes for §5.4); we report point estimates without confidence intervals. The headline $+0.021$ ensemble gain (§5.6) is therefore directional corroboration of a known effect, not a measured effect size; multi-fold/multi-seed validation with CIs is the most important piece of future work, which our deleted cloud infrastructure prevented us from adding before submission. (ii) **The leak claim is partly inferential.** We observe that the public ProtoSSM trains on the labelled soundscapes in-kernel, but only infer the SED's training scope from the 0.99-vs-0.95 optimism gap; the methodological conclusion holds under either label leakage or in-domain over-fit plus distribution shift. (iii) **Half-training cuts two ways.** The diverse ensemble was half-trained when a budget cap froze it. This understates its absolute ceiling (so the *deployment-wall* conclusion is conservative — a stronger ensemble would still time out), but it is also a genuine confound for the *substitution* result (§5.6): a fully-trained ensemble might have made the runtime-safe swap competitive, so we do not claim the swap result proves our members are intrinsically inferior. (iv) **The leaders are a conjecture.** Our characterization of the > 0.95 teams (§6) is unverified — their solutions are private. (v) **Scope.** All scores use this competition's masked macro-AUC on its specific data; the *platform-constraint* and *validation-trap* lessons should transfer qualitatively to similar code-competitions, but the exact thresholds will not.

8. What moved, and final standing

The only positive movement available to us was tracking the community's `TAX_SMOOTHING` jump to 0.950 and choosing robust final submissions. Recognizing the $\sim$890-team public tie, we selected two maximally-divergent 0.950 pipelines to hedge the public$\rightarrow$private shift. The shift proved mild and

uniform ($0.950 \rightarrow$ **0.942** across our picks); a small robustness signal favoured the more battle-tested calibration (the older EoS.8 base held 0.942 while the newer EoS.9 fell to 0.941). Final standing: **289 / 4243, bronze**.

9. Discussion

Three takeaways we believe are non-obvious and worth recording:

1. **Optimistic held-out can invert your conclusions.** When public components are fit on the only in-domain labelled data available for validation, offline comparisons systematically flatter them. Resource-constrained teams should treat such comparisons as *lower bounds* on their own models' relative quality, and prefer leak-free relative gates.
2. **Leak-aware offline gating.** Validating locally before submitting is standard practice — a local cross-validation split; the twist that mattered here is *what* to validate against. Because absolute held-out comparison to the baseline is biased (point 1), we gated on *relative* and *orthogonality* signals instead (ensemble-vs-best-single; per-class win counts against the public SED). This let us reach a decision on five directions within the budget a blind approach would spend on one, and we recommend the leak-aware variant specifically for code-competitions where the only in-domain labels were also used to fit the public baseline.
3. **In the foundation-model era the binding constraint is often the platform, not the idea.** The highest-value technique (pipeline ensembling) and a genuine $+0.021$ ensemble gain were both real and both undeployable under a 1 MB / 90-minute envelope. A public ceiling built from large shared pretrained components is not a wall of cleverness to out-think; it is a floor, and the door through it is pre-training compute a late, small team does not have.

Taken together these sharpen the question of where effort actually pays under foundation-model baselines and platform limits, and suggest that a useful contribution from a mid-pack team is a measured map of the terrain that winners' recipes leave unspoken.

10. Conclusion

We finished 289/4243 (bronze) at private 0.942, absorbed into the public 0.950 monoculture and unable to exceed it. The value of this note is not a score but a *negative result delivered with measurements*: a falsification of five improvement directions, identification of a held-out-validation trap, a reusable leak-aware local-CV gating methodology, and a platform-constraint explanation for why public foundation-model pipelines form such hard ceilings. We release all code, the cloud training/evaluation harness, and the gates openly, in the hope the map helps the next small team decide where to spend its five daily submissions.

Reproducibility & code

All model training (cloud GPU), the offline-gate harness (`compare_to_teacher`, `eval_sed_ensemble`, the probe and rare-taxon gates), soundscape pseudo-labeling, ONNX export, and kernel-build scripts are released at: https://github.com/Whyme-Labs/birdclef2026. The public-baseline components are the work of the Kaggle community and are credited to their original authors: Perch v2 (Google) [4]; the EoS / Light-ProtoSSM / distilled-SED / TAX_SMOOTHING lineage (nina2025, pilkwang, tuckerarrants, hideyukizushi, chaneyma, and others). This note's own contribution is the measurement framework and the negative-results map layered on top.

Acknowledgments

We thank the BirdCLEF+ 2026 organizers and the Kaggle community whose open notebooks constitute the baseline studied here.

Declaration on Generative AI

During the preparation of this work, the author(s) used Anthropic's Claude (via the Claude Code agent; see §4.2) in order to: assist with programming and debugging, run and parse experiments, review literature, and check grammar and spelling. All experiments were designed and configured under the authors' direction, all results were manually verified, and all scientific interpretations and conclusions reflect the authors' own judgment. After using these tools, the author(s) reviewed and edited the content as needed and take full responsibility for the publication's content.

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] S. Kahl, A. Navine, T. Denton, H. Klinck, P. Hart, H. Glotin, H. Goëau, W.-P. Vellinga, R. Planqué, A. Joly, 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, CEUR-WS.org, 2024.

[4] B. Ghani, T. Denton, S. Kahl, H. Klinck, Global birdsong embeddings enable superior transfer learning for bioacoustic classification, Scientific Reports 13 (2023) 22876. doi:10.1038/s41598-023-49989-z, perch / bird-vocalization-classifier embeddings; the competition uses the Perch v2 release distributed on Kaggle Models.

[5] A. Gu, T. Dao, Mamba: Linear-time sequence modeling with selective state spaces, arXiv preprint arXiv:2312.00752 (2023). arXiv:2312.00752.

[6] Q. Xie, M.-T. Luong, E. Hovy, Q. V. Le, Self-training with noisy student improves imagenet classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020. arXiv:1911.04252.

[7] L. Rauch, R. Schwinger, M. Wirth, R. Heinrich, D. Huseljic, M. Herde, J. Lange, S. Kahl, B. Sick, S. Tomforde, C. Scholz, Can masked autoencoders also listen to birds?, Transactions on Machine Learning Research (TMLR) (2025). Bird-MAE-Base checkpoint (DBD-research-group/Bird-MAE-Base).

[8] P.-Y. Huang, H. Xu, J. Li, A. Baevski, M. Auli, W. Galuba, F. Metze, C. Feichtenhofer, Masked autoencoders that listen, in: Advances in Neural Information Processing Systems (NeurIPS), 2022. arXiv:2207.06405.

[9] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, arXiv preprint arXiv:1503.02531 (2015). arXiv:1503.02531, neurIPS 2014 Deep Learning and Representation Learning Workshop.