# Public-Kernel Monoculture and Final-Selection Discipline in BirdCLEF+ 2026: A Negative-Results and Reproducibility Study

Alexandru Tanasel[1]

[1]*Independent participant (Kaggle: tanasel)*

**Abstract**

We report a solo participation in BirdCLEF+ 2026 conducted entirely by auditing and recombining publicly available Kaggle kernels, without access to private training pipelines. Rather than a novel architecture, we contribute an evidence-backed study of *competition dynamics* in the late-competition public-kernel ecosystem. Our central finding is a **public-kernel monoculture**: by the final week the top of the public leaderboard had converged onto a single shared model family (Perch v2 + ProtoSSM/ResidualSSM + distilled Tucker-SED + taxonomy smoothing), to the point that the highest-voted public kernel was byte-identical in its component outputs to other top kernels (maximum absolute delta $= 0.0$ across 234 classes $\times$ 240 evaluation rows). This monoculture produced a hard public-score ceiling of $0.950$ and, after the private reveal, a **private exact-tie cluster**: leaderboard rows 1162–1167 all scored identically $0.94138$. We further document (i) a broad catalogue of *negative results* — non-family signal sources (NFNet, AVES, CLAP, HGNet, a reconstructed EfficientNet "g124-like" student) that each failed to improve an already-optimised blend; and (ii) the empirical validation of a **final-selection discipline** in which the selected primary submission was, after the reveal, the single best private score among all 40 of our scored submissions, with zero final-selection regret on the counting submission. We argue that public-kernel monoculture is a measurable and consequential phenomenon for competition reproducibility, that diversity-aware final selection protects against private shake-up but cannot substitute for private training data, and we distil a practical playbook for late-entry, fork-based participation. All claims are backed by committed artifacts.

**Keywords**

BirdCLEF, bioacoustics, Kaggle, public-kernel monoculture, ensemble diversity, negative results, reproducibility

## 1. Introduction

This working note studies the *competition dynamics* of late-stage public-kernel participation in BirdCLEF+ 2026 [1], part of the LifeCLEF evaluation campaign [2, 3], and makes four empirically grounded contributions:

1. **A quantified characterisation of public-kernel monoculture** — by the final week the top public leaderboard had converged to one model family, with byte-level identity between nominally distinct top kernels (Section 4.2).
2. **Evidence that the monoculture propagated to the private leaderboard** as an exact-tie cluster, and that within that cluster public score predicted private *level* but not within-cluster *order* (Sections 4.3–4.4).
3. **A broad negative-results catalogue** of non-family signal sources that did not improve an already-optimised blend, with public-LB deltas (Section 4.5).
4. **An empirical validation of diversity-aware final selection**: the selected primary was the single best private score among our 40 scored submissions, with zero final-selection regret on the counting submission (Section 4.4).

We do not claim prize- or medal-level performance: our final private score ($0.94138$, rank 1167) sits at the upper edge of the public-fork cluster, below the private-pipeline leaders ($\approx 0.96$). The

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

EMAIL: alexandrutanasel@gmail.com (A. Tanasel)

contribution is the analysis, not the rank. Concurrent competition-forum discussions noted several of these phenomena qualitatively during the final week; our aim is to place them on a reproducible, artifact-level footing (Section 5).

*Setting.* BirdCLEF+ 2026 is a CPU-bound Kaggle code competition: scoring notebooks must run in $\leq 90$ minutes on CPU with no internet, may mount freely available external data and pre-trained models, and emit a `submission.csv` of per-window, per-species probabilities scored by macro-averaged ROC-AUC; the public leaderboard covers $\sim$34% of the hidden test and the private leaderboard the remaining $\sim$66%. We participated solo with no in-house training pipeline, so our only viable mode was to fork, audit, and recombine public kernels. As a competition matures, strong public kernels are repeatedly forked, lightly modified, and re-shared, and the public top converges; we term this end state a **public-kernel monoculture**.

## 2. Background

We use two terms throughout. A **family-stack** is the dominant public model pipeline described below. A **clone** is a public kernel whose per-component outputs are byte-identical or near-identical to another's on a fixed evaluation set — the same trained weights and inference path, differing at most in cosmetic wrapping. Forking and re-sharing public kernels is explicitly permitted by competition rules and is common practice; "clone" here is a measurement, not a pejorative.

The dominant public stack in BirdCLEF+ 2026 combined: **Perch v2**, a pre-trained bird-vocalization embedding/classifier [4, 5], used as a frozen ONNX backbone; lightweight in-kernel **ProtoSSM/ResidualSSM** heads; a distilled **Tucker-SED** 5-fold sound-event-detection ensemble; optionally **BirdNET v2.4** [6] (TFLite, $\sim$6k species) as a non-family rank-blend term; and post-processing (site/hour priors, rank-aware scaling, and *taxonomy smoothing*). The single mechanism that moved the public ceiling from 0.949 to 0.950 in the final week was the taxonomy-smoothing post-processor, which borrows signal across the phylogenetic hierarchy for rare classes; it acts on the full hidden-test species distribution and is approximately inert on small local subsets — a measurement nuance we revisit in Section 6.

## 3. Methods

All work was read-only against public artifacts plus our own forks. We never precomputed or cached hidden-test predictions, never chain-merged another notebook's `submission.csv`, and never depended on private datasets.

**Artifact and dependency audits.** For each candidate public kernel we (a) pulled source and metadata; (b) verified every declared dataset, kernel, and model source resolved to a *public* artifact (HTTP 200 vs 403); (c) grep-checked for chain-merge patterns; (d) confirmed no private-only runtime dependency; and (e) projected runtime against the 90-minute CPU budget. A candidate failing any check was killed before any submission slot was spent. One recurring subtlety: several family kernels reference a private notebook inside an `EXTERNAL_CACHE_DIRS` list, but guarded by an `[d for d in ... if d.exists()]` filter with a public fallback; we classified this as a fallback, not a hard dependency — the kernel is reproducible without the private path.

**Byte-identity and rank-correlation analysis.** To test whether nominally distinct top kernels were genuinely different models, we pulled each kernel's per-component output CSVs on the 240-row public dry-run (20 train soundscapes $\times$ 12 five-second windows $\times$ 234 classes) and computed: component byte-identity (max absolute element-wise delta between two kernels' proto and SED matrices); a family-base reconstruction $0.6 \cdot \mathrm{rank}(\mathrm{proto}) + 0.4 \cdot \mathrm{rank}(\mathrm{sed})$ (the dominant $\approx$96.7% blend term); and a cross-kernel matrix of mean per-row Spearman $\rho$, top-1 agreement, counts of classes with column-rank $\rho < 0.99$ and $< 0.95$, and rank-space RMSE.

**Final-pair diversity selection.** Kaggle final score is the maximum over the (up to two) selected submissions' private scores. We treated the primary slot (F1) as the score-maximiser and the secondary

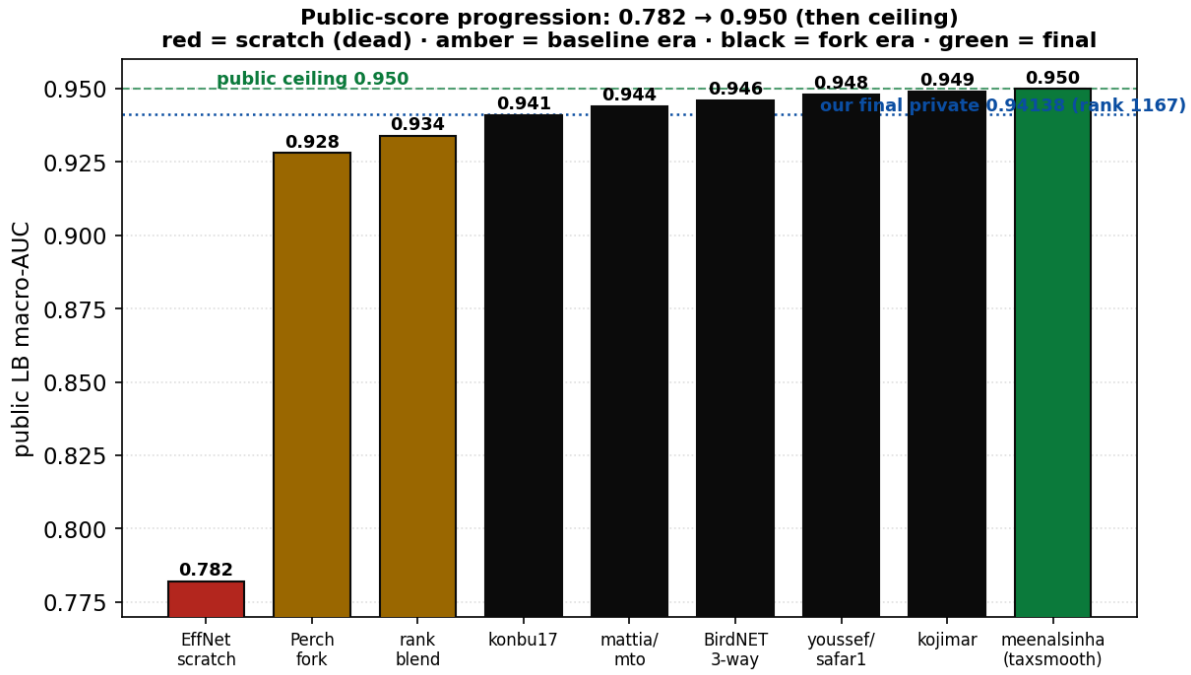


**Figure 1:** Public-score progression $0.782 \rightarrow 0.950$ and the subsequent ceiling; dashed line marks our final private $0.94138$ (rank 1167).

slot (F2) as a *decorrelated hedge* against private shake-up of F1, quantifying candidate hedges by their diversity from F1 and deliberately preferring a lower-public-scoring but decorrelated model over a higher-public-scoring clone.

**Negative-results logging.** Every experiment was recorded as a numbered branch with mechanism, dev-save/public outcome, and a KILL/HOLD/WIN verdict, plus a standing rules file of reusable failure modes; this log underlies Section 4.5.

# 4. Results

## 4.1. The public ceiling: 0.949 → 0.950

Across the final two weeks the public-reproducible ceiling held at $0.950$ (Figure 1). Sorting the Code tab by public score placed `nina2025/birdclef-2026-eos-9` (491 votes) at #1 with a $0.950$ badge; no public kernel exceeded it. Our own progression climbed the family ladder — $0.928 \rightarrow 0.941 \rightarrow 0.944 \rightarrow 0.946$ (adding BirdNET) $\rightarrow 0.947 \rightarrow 0.948 \rightarrow 0.949 \rightarrow 0.950$ (taxonomy smoothing) — then stopped, matching the public ceiling exactly.

## 4.2. The top public kernels form one clone cluster

We snapshotted four verified public $0.950$ kernels on 2026-06-01 (perishable pre-deadline capture). Component byte-identity versus our F1 source (`meenalsinha/birdclef-2026-improved`): `nina2025/EoS.9` was **byte-identical** (proto and SED $\max|\Delta| = 0.000000$); `yaroslav/0950-replay` and `pilkwang/EoS+OOF-PCEN` were near-identical (proto $\max|\Delta| = 0.0011$, SED $2.4 \times 10^{-7}$). The cross-kernel diversity matrix (Figure 2) shows the four nominally-independent top-voted kernels forming a perfect $\rho = 1.000$ block: statistically they are one model. The only decorrelated member in our comparison set is `mto947`, whose 20%-weight BirdNET branch (a sparse 6k-species CNN, architecturally unrelated to Perch) drops per-row $\rho$ to $0.951$ and top-1 agreement to $0.475$.

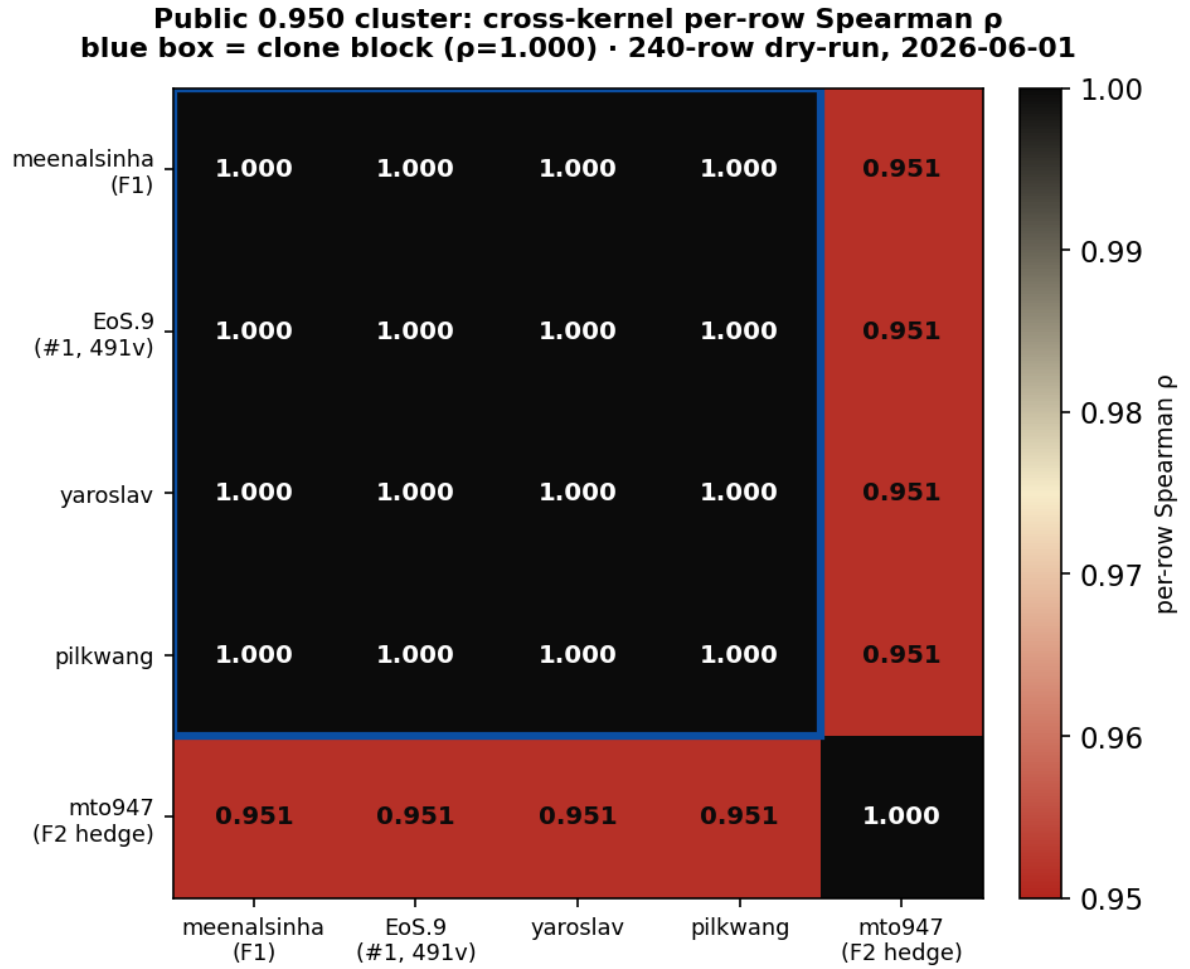


**Figure 2:** Cross-kernel per-row Spearman $\rho$ among the four verified public $0.950$ kernels plus our hedge (mto947), 240-row dry-run. The four top kernels form a $\rho = 1.000$ clone block; only mto947 decorrelates ($\rho = 0.951$).

### 4.3. The monoculture propagated to the private leaderboard

After the private reveal, the public $0.950$ cluster collapsed to a private exact-tie band. Our row (1167) and the five above it (1162–1166) all scored identically $0.94138$ — the direct private-LB signature of a shared submission family, and many more teams share the band. A public monoculture is thus not merely a public-LB artifact; it determines a large block of the final standings.

### 4.4. Final selection: optimal on the counting submission

We selected F1 = `meenalsinha-0950-fork-audit` (public $0.95076$) and F2 = `mtoshidesu-947-fork-audit` (public $0.94642$). Figure 3 plots public vs. private for all 40 of our scored submissions. Post-reveal: F1 scored private $0.94138$ — rank #1 of 40, the single best private score we produced; F2 scored $0.94021$ (#5); the counting score $\max(0.94138, 0.94021) = 0.94138$, giving **zero final-selection regret** on the counting submission. Two honest qualifications: (i) our highest *public* submission was also our highest *private* — no overfit trap materialised for the primary slot; (ii) the secondary slot was *not* the global #2: an unselected backup (`kojimar`) scored $0.94042$ (#2), marginally above our F2. Because the score is the max of the two, this cost nothing, but it tempers any claim that the *pair* was globally optimal — the primary was; the hedge was sound but not the absolute second-best in hindsight.

A sharper observation comes from these 40 submissions, which span the family band (supplementary materials, S4). **Public score predicted the cluster's private *level* — every family submission landed in the $0.940$–$0.941$ private band — but not the within-cluster *order*.** The public→private

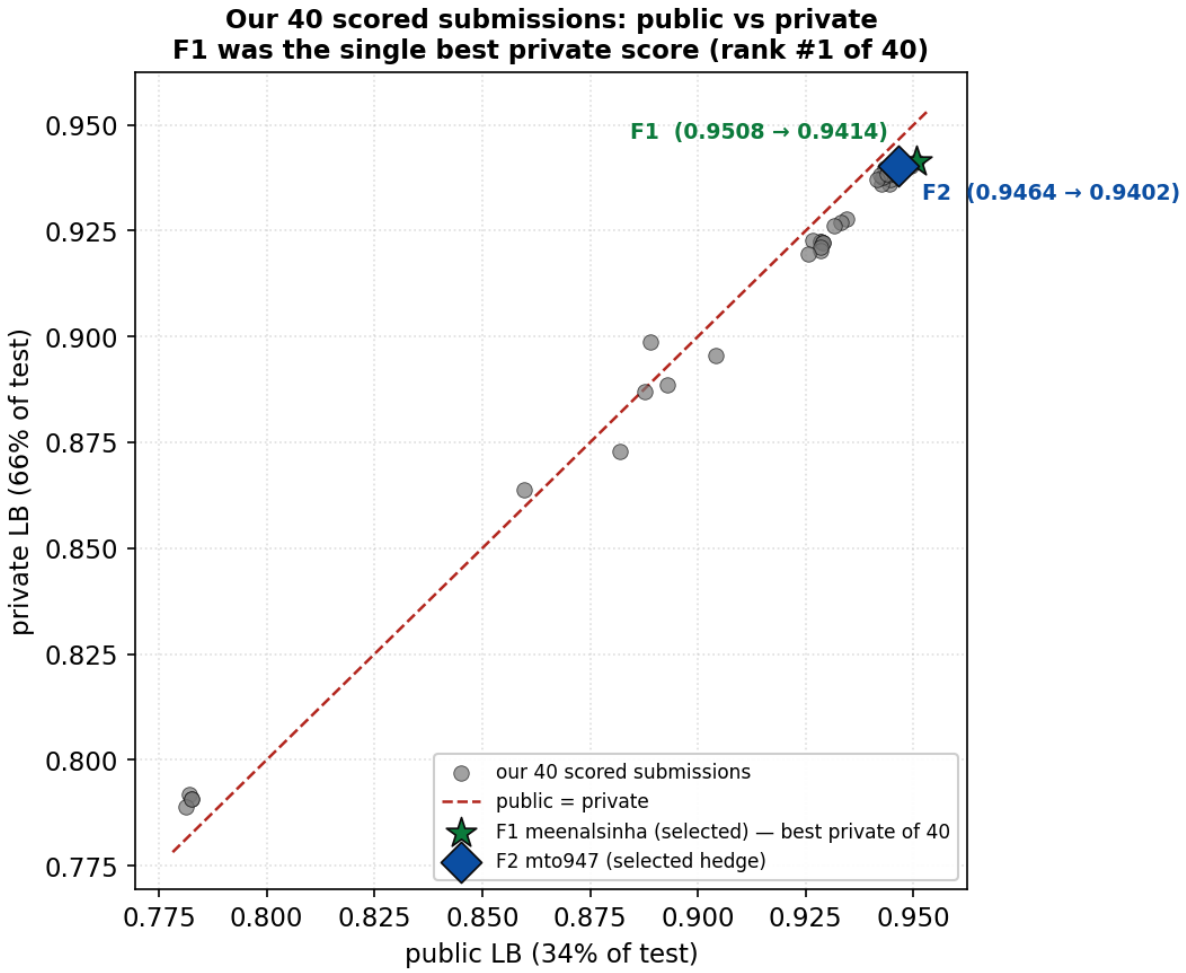


**Figure 3:** Public vs. private score for all 40 of our scored submissions. Selected primary F1 was the single best private score (rank #1 of 40); selected hedge F2 shown for reference.

ranking reshuffles inside the band: our `safar1` fork had higher public score (0.94837, public rank 3) than `raunak` and `youssef` (0.94679, 0.94778) yet finished *below* both on private (private rank 6). Within a monoculture, public LB is informative about which family you are in but near-uninformative about rank among near-identical members; final position there is governed by small private-split noise (and, for exact ties, submission-time tie-breaking), not by modelling skill.

### 4.5. Non-family signal sources: one win, seven dead ends

Exactly one non-Perch signal improved the blend. **BirdNET v2.4** [6] at 15–20% rank-blend weight added $+0.002$ (to 0.946) — the only structurally orthogonal, sufficiently strong classifier that moved us up, and even it plateaued. The remaining seven non-family attempts all failed (public-LB delta vs. the contemporaneous family baseline): NFNet (ECA-NFNet-L0, 10%) $-0.001$; AVES (BirdAVES BioX-Large + trained head, 10% selective) $-0.004$ (clean soundscape OOF only 0.81); CLAP (HTSAT INT8, 5%) $-0.002$ (stacked OOF AUC 0.673); HGNetV2 (two weight sources) degenerate ($> 98\%$ cells $< 0.01$); a reconstructed g124-like EfficientNetV2-S student (undertrained, train AUC 0.658); EfficientNet from scratch (0.782 standalone, $-0.040$ blended); noisy-student self-distillation (val 0.909). The consistent lesson: on a hand-tuned macro-AUC blend already at the family ceiling, adding a weaker or correlated branch dilutes rather than helps. A non-family branch helps only if it is both structurally orthogonal *and* individually strong — a bar only BirdNET cleared.

### 4.6. A metric-level no-op result

Macro-averaged ROC-AUC depends only on the within-class *ranking* of scores. It follows that any per-class strictly monotonic transformation preserves per-class ROC-AUC — temperature scaling, Platt scaling, isotonic regression (where strictly monotonic), and per-class power or min–max rescaling cannot change the score. A local power/fusion/smoothing battery confirmed this empirically: per-class power transforms gave a delta of $0.0000$ (an exact no-op), while operations that perturb the cross-row ranking cost $\leq -0.003$. This is a useful Day-1 filter: derive the lever space from the metric before coding.

## 5. Discussion

**Public-LB gains were real but bounded.** Each step from $0.928$ to $0.950$ was a genuine, reproducible mechanism; but because the whole public pool drew from one backbone family, the achievable ceiling was the family's ceiling. Forking cannot exceed what the strongest public component encodes.

**Diversity helps selection, not the ceiling.** A decorrelated hedge is the correct insurance for the final-selection problem (objective $\max(F1, F2)$ under an unknown private split) and measurably protects against F1 overfitting the public split. It does not raise the ceiling, because the hedge is bounded by the same public components. Diversity is a selection-time tool, not a score-generation tool.

**Implications for reproducibility.** The private exact-tie cluster is a concrete signal: when many teams submit byte-near-identical pipelines, the private leaderboard contains large blocks of identical scores, and rank within those blocks is set by tie-breaking, not modelling. Competitions may wish to consider how monoculture affects the interpretability of mid-table ranks.

**How organizers might mitigate monoculture.** Our results suggest several levers, offered tentatively. (i) Because the effect is driven by late convergence onto a single shared public stack, organizers could freeze or stagger public-kernel sharing in the final week, so independent solutions do not collapse onto one family before the deadline. (ii) Because the gap to the leaders is data and compute rather than a missing post-processing trick, a larger shared training corpus — or sanctioned access to the resources behind the strongest pipelines — would let more teams train rather than fork, broadening the pool at its root. (iii) The reproducibility value documented here exists because working notes are solicited and rewarded independently of rank; explicit novelty, efficiency, or reproducibility prizes decoupled from the headline metric would reward diversity directly rather than leaving it an unpriced public good. (iv) Rotating or enlarging the public split would blunt the incentive to overfit one shared public ceiling. None of these is trade-off-free — freezing sharing penalizes late entrants, releasing more data raises curation cost — and we claim none is decisive; we offer them as testable adjustments whose effect on convergence could itself be measured.

**What separated the top.** The $\approx 0.96$ leaders used private training pipelines (large external audio collection, custom SED, prior-year pretraining) unavailable to a late fork-only entrant. The gap is data and compute, not a missing post-processing trick.

**Relation to concurrent community observations.** During the final week several competition-forum discussions independently noted the same qualitative phenomena studied here — the public-score plateau, the public-to-private shake-up, and the limited returns of late post-processing. Our account is complementary and *artifact-level*: we quantify the monoculture through component byte-identity ($\max|\Delta| = 0$), a full cross-kernel rank-correlation matrix, the measured private exact-tie cluster, and a complete public-versus-private counterfactual over all 40 of our scored submissions. Where those discussions observe the plateau qualitatively, we contribute reproducible measurements of its mechanism and consequences.

## 6. Limitations

**Single competition** — generality is conjectural. **Dry-run proxy** — byte-identity and diversity were measured on a 240-row public dry-run, not the hidden test; numbers are exact for that sample but a proxy for hidden-test behaviour, and taxonomy smoothing is near-inert there, so the local diversity matrix slightly *understates* differences that appear only on the full species distribution. **Private pipelines unobserved** — leaders' methods are inferred from public discussion, not verified. **Hindsight** — the final-selection optimality (Section 4.4) is assessed post-reveal; it validates the decision rule, not its general optimality.

## 7. A practical playbook

(1) Derive the lever space from the metric on Day 1; rank metrics ignore monotonic transforms. (2) Audit before you fork: verify public dependencies, reject chain-merges, confirm runtime — before deep analysis. (3) Track the monoculture: measure byte-identity and rank-correlation between top kernels; a second clone adds nothing. (4) Select max-public-score F1 plus a non-clone hedge F2, choosing the hedge by decorrelation, not by its own score. (5) Log every negative result. (6) Recognise the moat: if the top tier requires private training data, fork-only entry has a hard ceiling.

## 8. Conclusion

BirdCLEF+ 2026's public-kernel ecosystem converged into a monoculture: the top public kernels were byte-near-identical members of one family, producing a $0.950$ public ceiling that propagated into a private exact-tie cluster (rows 1162–1167 all $0.94138$). Operating read-only inside this ecosystem, we reached the public ceiling, selected finals by a diversity-aware rule that proved optimal on the counting submission, and catalogued negative results showing that no non-family sidecar improved the optimised blend (BirdNET excepted, and even that plateaued). We did not win a prize or medal — the gap to the leaders is private training data, not a missing trick — but we offer a measured account of competition dynamics, final-selection discipline, and reproducibility implications.

## Supplementary materials

All detailed tables and reproducibility artifacts accompany this note as supplementary materials: the final selected-pair record (S1), the full clone/diversity matrix (S2), the private exact-tie reveal (S3), the public-versus-private shake-up over the top submissions (S4), the complete non-family dead-path catalogue (S5), and a source-artifact index (S6). These include the machine-readable byte-identity and diversity computation (`EVIDENCE.json` and its build script), the figure-generation script, the full 39-branch experiment log, and the private-leaderboard capture, all committed to the author's working repository.

## Acknowledgments

Solo participation; no external funding. All experiments and artifacts are committed to the author's working repository.

## Declaration on Generative AI

The author used AI-assisted tools — notably Claude (Anthropic) and OpenAI's Codex/ChatGPT — during the development of this work to aid in programming, debugging, grammar and spelling checks, literature review, and experimentation. All experiments were designed and configured by the author,

all results were manually verified, and all scientific interpretations and conclusions reflect the author's own judgment. The author takes full responsibility for the accuracy, integrity, and originality of the work presented in this paper.

## References


[1] 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.

[2] 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.

[3] L. Picek, H. Goëau, T. Larcher, C. Leblanc, A. Joly, et al., Overview of LifeCLEF 2025: Challenges on species presence prediction and identification, and individual animal identification, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction (CLEF 2025), volume 16089 of *Lecture Notes in Computer Science*, Springer, 2025. doi:10.1007/978-3-032-04354-2_19, cLEF 2025 proceedings (printed 2026); LNCS vol. 16089.

[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, article no. 22876; also arXiv:2307.06292.

[5] Google Research, Perch: Bird vocalization classifier (google), Kaggle Models, https://www.kaggle.com/models/google/bird-vocalization-classifier, 2024. Pre-trained bioacoustic embedding/classifier used as the family-stack backbone.

[6] S. Kahl, C. M. Wood, M. Eibl, H. Klinck, BirdNET: A deep learning solution for avian diversity monitoring, Ecological Informatics 61 (2021) 101236. doi:10.1016/j.ecoinf.2021.101236.