MARINEMAMMALBENCH: SAMPLE RATE AND EMBEDDING GRID AS LOGGED VARIABLES IN A BENCHMARK FOR MARINE MAMMAL SELF-SUPERVISED AUDIO ENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-supervised (SSL) audio encoders are now routinely reused far from the speech and general audio they were trained on. Two preprocessing choices then matter a great deal and are rarely reported: the input sample rate, and therefore the acoustic bandwidth the encoder retains, and the temporal grid on which it emits embeddings. We treat both as logged, swept experimental variables and study them in marine mammal passive acoustic monitoring, where the discriminative energy runs from killer-whale call types below 8 kHz to sperm-whale clicks above 20 kHz. We propose *MarineMammalBench*: a two-track protocol (in-domain SSL pretraining on a curated $\sim$2.17M-hour unlabelled hydrophone pool; frozen-probe screening of public encoders under a matched protocol), a shared task ladder, and four mandatory controls, including a grid–tolerance coverage diagnostic. First-generation results give five findings. In a fixed 32 kHz sperm-whale click detector, restricting bandwidth to 3.5 kHz reduces event F1 from 0.7246 to 0.3999 at 20 ms while detector weights and the embedding grid remain unchanged. The historical 32–8 kHz F1 gap grows $2.8\times$ when the matching collar tightens from 20 to 2.5 ms, so the temporal evaluation rule sets the size of the effect. For killer-whale call types the bandwidth gain is small (0.015 macro-F1 from the band above 8 kHz), so bandwidth has to be evaluated per task. On the official 31-class BEANS Watkins split, frozen marine ANIMAL2VEC reaches 92.04% accuracy, ahead of PERCH 2.0 (89.97%) and WHISPER-tiny (87.91%); validation-selected readout raises Whisper accuracy from 75.22% at its final representation. On HICEAS minke-whale detection, AVES-bio instead leads the linear comparison with AP 0.5044, showing that encoder ranking depends on the task. On the prior system's presence test set, whether the model beats the human annotators depends on the metric and on the precision floor. The controls make bandwidth, temporal evaluation and uncertainty explicit parts of marine SSL benchmarking.

1 INTRODUCTION

Self-supervised audio encoders have become the default recipe for recognition in low-resource acoustic domains: pretrain once on unlabelled audio, then attach small probes per task. When these encoders move from speech and general audio into new domains, a choice that speech benchmarks never force becomes important: the input sample rate, and with it the acoustic bandwidth that survives. Marine mammal passive acoustic monitoring (PAM) is where this matters most. Hydrophone deployments capture millions of hours at 48 kHz and above, spanning a discriminative range far wider than speech, and manual annotation is slow and expensive, so learned detectors increasingly triage, filter and classify the archives (Gibb et al., 2019; Stowell, 2022).

Many widely used audio SSL models operate at 16 kHz, capping retained bandwidth at $\sim$8 kHz. Odontocete clicks, whistles and burst pulses extend well above that, so an encoder may perform poorly not because its objective is weak but because the evidence was removed before the model saw the input. A second variable comes with the first: input rate also fixes the temporal grid on which frame embeddings are emitted, and event-level scores depend on that grid through the matching

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

tolerance. Neither quantity is isolated by existing bioacoustic benchmarks (Hagiwara et al., 2023; Hamer et al., 2023; Rauch et al., 2024; Earth Species Project, 2025) nor by the closest marine SSL pipeline, DORI (Nestor et al., 2026), which fixes one operating point (32 kHz through a 16 kHz front end, 1 kHz high-pass) and never sweeps it.

Contributions. (i) A benchmark and training protocol in which sample rate, resampling strategy, retained bandwidth and embedding grid are logged, swept variables (Sec. 3), with a primary in-domain SSL training track over ∼2.17M unlabelled hours and a frozen-probe baseline track under one matched protocol. (ii) Four mandatory controls: a grid–tolerance coverage diagnostic, an annotation-geometry baseline, a published human band and a stated split rule (Sec. 3.4), with empirical evaluation uncertainty reported alongside (Sec. 5.2). (iii) First-generation results (Sec. 5): a fixed-detector intervention shows that click detection depends on bandwidth at unchanged input rate and embedding grid; a collar sweep shows that the size of the rate effect depends on the matching tolerance; the call-type comparison shows a much smaller bandwidth gain; and official BEANS Watkins/HICEAS evaluation shows task-dependent encoder rankings and substantial gains from validation-selected feature readout. (iv) Transparent evaluation: we report model selection, limited run replication and known confounds, with paired file-bootstrap uncertainty for the bandwidth intervention.

2 Related Work

BEANS (Hagiwara et al., 2023), BIRB (Hamer et al., 2023) and BirdSet (Rauch et al., 2024) are the closest precedents for broad bioacoustic benchmarking; none isolates sample rate or retained bandwidth as an axis. Earth Species Project (2025) compare pretrained encoders on bioacoustic transfer, mainly for birds, and likewise hold bandwidth fixed. AVES (Hagiwara, 2023), Perch 2.0 (van Merriënboer et al., 2025), animal2vec (Schäfer-Zimmermann et al., 2024) and NatureLM-audio (Robinson et al., 2024) represent the animal-domain encoder line; speech and general-audio encoders (wav2vec 2.0, WavLM, BEATs, AudioMAE, Whisper) are the usual transfer baselines (Baevski et al., 2020; Chen et al., 2022; 2023; Huang et al., 2022; Radford et al., 2023). Frozen-feature probing follows the HEAR convention (Turian et al., 2022). Bat bioacoustics has long treated ultrasonic bandwidth as central (Mac Aodha et al., 2018); marine PAM spans killer-whale call types (Ford, 1987; Filatova et al., 2017; Bergler et al., 2019; 2022), sperm-whale clicks (Bermant et al., 2019; Gubnitsky and Diamant, 2024; Gubnitsky et al., 2024) and DCLDE-style detector evaluations (DCLDE, 2024).

DORI. The closest prior system (Nestor et al., 2026) releases ∼5,300 h of Northeast Pacific audio with detection, species and ecotype labels and reports that a frozen Whisper-tiny embedding with logistic regression is competitive with specialised detectors. Its front end matters here: audio is resampled to 32 kHz "to maintain a Nyquist frequency of 16 kHz", high-passed at 1 kHz, and fed through Whisper-tiny's default 16 kHz mel front end, so retained bandwidth is 1–16 kHz. DORI documents that trade-off but does not vary it; backbone, objective and corpus are likewise fixed. We turn each into a logged axis and include Whisper-tiny as a frozen baseline so DORI's backbone is scored on the same ladder. We take its operating point as one sample of the space we sweep, not as an error to correct.

3 Benchmark Design

3.1 Two tracks, one task ladder

The *primary track* trains and retrains in-domain marine SSL encoders on the unlabelled archive pool, varying objective, architecture, model size, input rate and corpus, with a sound-filter retraining loop (generation 1 → filter → generation 2). The *baseline track* evaluates public encoders frozen: an encoder f_m maps clip x_i to $z_i = f_m(x_i)$ and a per-task probe p_t gives $\hat{y}_i = p_t(z_i)$; the probe protocol is identical in both tracks, so in-domain gains are read off directly. Both tracks feed one task ladder (Table 1, Figure 1).

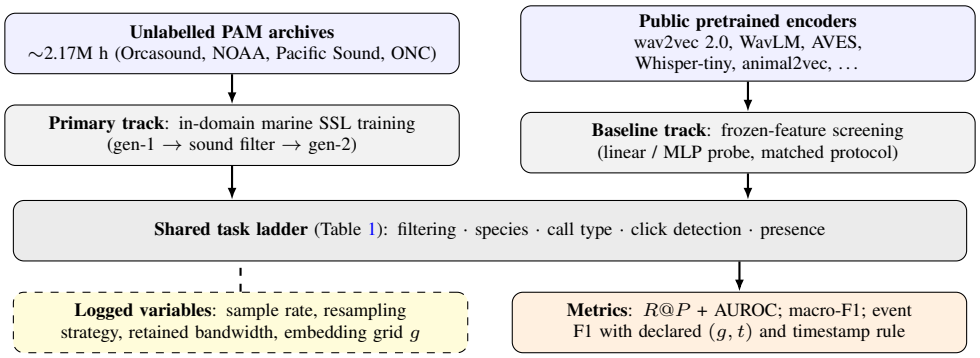


Figure 1: Two-track design. The primary track trains in-domain marine SSL encoders on unlabelled archives with a sound-filter retraining loop; the baseline track screens public encoders frozen. Both feed one task ladder under one probe protocol.

Table 1: Benchmark task ladder. "Run" marks rungs with results in Sec. 5; "specified" marks rungs fixed here but deferred; "partial" marks a rung with a diagnostic but no retraining arm. Every rung reports the controls of Sec. 3.4.

Rung	Question	Primary metric	Degenerate baseline	Status
T1 Sound filtering	signal vs. noise/artifact triage	macro-F1	majority class; broadband energy	Run
T2 Species ID	31-way BEANS Watkins species	accuracy; macro-F1	majority class; clip duration	Run
T3 Call type	killer-whale call type, 11 types + noise (12-way)	macro-F1	majority class	Run
T4 Click detection	sperm-whale echolocation events	event F1 at common tolerance; frame AUROC	energy threshold; click-rate prior	Run
T5 Presence	DORI 5-min files; HICEAS 10-s windows	$R@P$/AUROC (DORI); AP (HICEAS)	prevalence; **human band** (DORI)	Run
T6 Cross-deployment	transfer to an unseen hydrophone	$R@P$, worst-fold	site-identity probe	Specified
T7 Range/SNR	where detection fails with distance	$R@P$ per SNR stratum	prior on encounter rate	Specified
T8 Label integrity	is a propagated label exchangeable?	label-provenance AUROC	metadata-only predictor	Partial
T9 Cost	joules and latency at fixed recall	J/h at matched $R@P$	smallest encoder	Specified

3.2 Sample rate, resampling and retained bandwidth as logged variables

Every experiment logs four quantities: the *original recording rate*; the *model input rate*; the *resampling strategy* (target rate, anti-alias filter, controlled downsampling or band-limited control); and the *retained bandwidth* that survives. A 96 kHz recording resampled to 16 kHz for WAV2VEC 2.0 retains 0–8 kHz; the same clip at 32 kHz for PERCH 2.0 retains 0–16 kHz; a native log-mel pipeline retains 0–48 kHz (worked example in Appendix A). For each high-rate dataset we run, where feasible, model-default resampling, controlled downsampling to $\{8, 16, 32\}$ kHz plus a native reference, and band-limited controls. A fixed input rate imposed by a pretrained model is treated as part of the model. Event-level tables additionally log the *physical embedding interval* g and the *matching tolerance* t.

3.3 Task ladder

T1–T4 ask whether an encoder's features linearly separate the classes of interest under a matched frozen probe; T4 is the bandwidth stress test: sperm-whale clicks are the one family in the ladder whose discriminative energy clearly extends past 8 kHz. T5 includes DORI's released ONC benchmark (385 five-minute files, expert-labelled, with three amateur annotator columns), the one rung with a *published human band*. BEANS HICEAS adds minke-whale presence in 10-second windows under its official split. T6–T9 measure deployment behaviour (unseen hydrophone, range/SNR strata, label provenance, cost at fixed recall) and are specified here (Appendix C) for follow-up work.

3.4 MANDATORY CONTROLS

Grid–tolerance coverage. Every event-level number declares the embedding interval g, matching tolerance t and timestamp extraction rule. For timestamps constrained to a lattice of spacing h, with ground-truth phases uniformly distributed modulo h, $q(h,t) = \min(1, 2t/h)$ is the expected fraction within tolerance of a lattice point. This is a geometric coverage figure, not an F1 ceiling; event collisions can reduce recall further. The timestamp lattice may differ from the embedding grid, so $q_g = \min(1, 2t/g)$ is reported as nominal encoder-grid coverage. High frame AUROC with low event F1 points to thresholding, event extraction or localisation, without saying which. **Annotation-geometry baseline.** Where a window is cut from an annotation box, a baseline trained on the box geometry alone (duration and frequency bounds, no audio) is required. On the DCLDE ecotype labels under the 4-train/5-test protocol of prior work, geometry alone reaches 0.398 top-1 and duration alone 0.353, against a 0.215 majority floor and a 0.841 ceiling; median box bandwidth per ecotype spans 288 Hz to 6,198 Hz, an annotation convention rather than a biological signature. **Human band.** Where several annotators exist, their agreement with the reference is reported beside model scores: on T5 the three amateur annotators span positive-class F1 0.62–0.81 (κ 0.23–0.71) at operating points from $P{=}0.99/R{=}0.69$ to $P{=}0.51/R{=}0.79$ on identical audio. **Split rule.** The benchmark requires recording, session, deployment or encounter-level splits, with the unit stated per rung. The filtering and GMWM analyses state where their splits fall short of this rule.

3.5 DETECTION METRIC AND PROTOCOL

For event-level detection the primary number is *recall at fixed precision* ($R@P{=}0.5$, $R@P{=}0.8$), alongside AUROC, average precision and event F1 with a collar sweep (Mesaros et al., 2016; Bilen et al., 2020); in PAM review a missed call is costlier than a discardable false positive. Classification uses accuracy for official BEANS Watkins, with macro-F1 secondary; other classification tasks retain macro-F1. BEANS HICEAS window presence uses AP. Embeddings are extracted once per encoder–dataset–rate condition and cached; linear and shallow-MLP probes are fitted with hyperparameters chosen on validation. Full split definitions and probe details are in Appendix D.

4 DATASETS AND ENCODERS

The labelled v1 corpus comprises a manual sound-filtering set (∼15 h), Watkins MMSDB (∼5 h, up to 166 kHz), Orcasound labelled audio (∼1,900 h), the Avacha-Gulf killer-whale corpus (Filatova et al., 2017) (48 kHz; 11 catalogued call types plus an ambient-noise class, and 17 annotated field recordings), BEANS HICEAS minke-whale recordings, Voices in the Sea, DCLDE 2022/2024, and a Dominica sperm-whale click set with event-level times of arrival. The unlabelled pool (Orcasound archive, NOAA/NCEI, Pacific Sound, ONC) totals ∼2.17M hours, two orders of magnitude more than DORI's curated corpus; we do not assume all hours contain vocalisations. The first pretraining corpora drawn from it hold 3,090 h at 8 kHz and 3,542 h at 16 kHz (measured from manifests). Encoders screened frozen are WAV2VEC 2.0, WAVLM, WHISPER-tiny, AVES, ANIMAL2VEC, PERCH 2.0 and non-SSL log-mel/random-CNN controls; in-domain variants are ANIMAL2VEC-style encoders trained on the pool at $8/16/32$ kHz, and AVES-HuBERT with domain-adaptive continued pretraining (DAPT, Gururangan et al., 2020). Full tables are in Appendix B. We also test the released Google Multispecies Whale Model (GMWM) as a frozen feature extractor for Avacha K-call recognition (Harrell, 2024; Google Research, 2024). Its short-clip classifier is evaluated on continuous recordings in Sec. 5.6.

5 FIRST-GENERATION RESULTS

Status and scope. Watkins and HICEAS use the official BEANS release with validation-selected readout and regularization, frozen before test extraction. Native frontends, model contexts and pretraining histories differ, so the frozen comparison measures transferred representations rather than an isolated pretraining objective or bandwidth effect. Probe seeds and evaluation bootstraps do not quantify pretraining-seed variation; public benchmark membership does not establish source or pretraining disjointness. The retained K-class/filter screen uses historical checkpoint and layer selection on reported scores, with unequal search budgets across families; its estimates do not share the new

Table 2: Official BEANS frozen-feature comparison with a common balanced linear probe. Watkins uses 31 classes and 1,017/339/339 train/validation/test clips; accuracy and macro-F1 are percentages. HICEAS detects minke-whale presence in 10 s windows with 5 s hop; AP is the pinned BEANS rank metric on 1,485 test windows. Readout and regularization are selected on validation accuracy (Watkins) or AP (HICEAS), before test extraction.

Encoder / features	Watkins 3 s Acc.	Watkins 3 s Macro-F1	Watkins 5 s Acc.	HICEAS 10 s Linear AP
Original AVES-bio	84.66	85.23	84.37	**0.5044**
WHISPER-tiny	87.91	88.39	87.91	0.4988
ANIMAL2VEC16 BEST129k	**92.04**	**92.70**	**91.74**	0.3515
PERCH 2.0	89.97	90.28	89.38	0.4735
Log-mel control	82.60	83.09	85.25	0.3510
Random CNN features	80.53	80.63	81.42	0.2440

BEANS selection protocol. Historical classification and click-rate variants also change physical window length.

5.1 FROZEN-PROBE ENCODER COMPARISON

Marine animal2vec leads the official species comparison. With a common trained linear probe and validation selection, animal2vec16 BEST129k reaches 92.04% accuracy and 92.70% macro-F1 on Watkins (Table 2). Perch 2.0 reaches 89.97% accuracy, Whisper-tiny 87.91%, and original AVES-bio 84.66%; log-mel and random CNN features reach 82.60% and 80.53%. Native frontends and pretraining histories differ, so this is a comparison of transferred representations.

Supervised CNN control. A small CNN trained on Watkins labels without SSL pretraining uses the same convolutional architecture as the random-feature control. It reaches $82.99 \pm 2.07\%$ accuracy with 3 s input and $80.24 \pm 2.99\%$ with 5 s input (mean $\pm$ sample SD over three seeds). This provides a supervised reference for assessing transfer from pretrained representations.

Feature readout changes transfer performance. Selecting readout on validation raises Whisper accuracy from 75.22% with native-final features to 87.91%. The selected AVES and Whisper readouts are early; animal2vec selects an intermediate block. Extending observed audio from 3 to 5 s does not consistently improve the pretrained encoders.

Encoder ranking depends on the task and head. In the linear HICEAS comparison, original AVES-bio leads with AP 0.5044, followed by Whisper-tiny (0.4988), Perch 2.0 (0.4735) and animal2vec16 (0.3515). The strongest species representation therefore does not lead window detection. With the secondary MLP head, Perch reaches AP 0.5411 ± 0.0256 (mean and sample SD over three seeds). The AVES lead applies to the linear comparison; the choice of head can change the ranking.

Historical call-type and filtering comparisons. Table 3 retains the tasks not repeated in the new BEANS comparison. Whisper-tiny leads K-class (90.67 macro-F1 versus 89.66 for AVES-DAPT and 66.92 for the animal2vec baseline), while AVES-DAPT leads filtering (99.10). The selected Modern2026 checkpoints score below the animal2vec baseline on both tasks; these selected runs do not isolate the effect of positional encoding. Original AVES-bio in Table 2 and historical AVES-HuBERT DAPT are distinct checkpoints.

5.2 MEASUREMENT UNCERTAINTY

The BEANS linear results are test point estimates; no recording-group confidence interval was computed. MLP seed SD measures head-training variation under a fixed split and pretrained checkpoint. Field comparisons use paired recording-bootstrap intervals, and the fixed-detector bandwidth experiment uses paired file bootstrap (Appendix F.1). These procedures quantify different sources of variation; none measures pretraining-seed uncertainty.

Table 3: Historical frozen-probe macro-F1 (%) for tasks not repeated in BEANS/E1: Avacha K-class (11 call types plus noise) and binary signal/noise filtering (500/500 clips). Original checkpoint and per-task layer choices are retained, including the earlier three-task checkpoint-selection criterion; selection used reported scores, so these are optimistic historical estimates. No mean across the historical and new tasks is reported. Physical/logical rate is shown as 32/16; windows vary with rate. Modern2026 changes positional encoding, normalization and attention; multires changes the frontend. Protocols and grouping limitations are in Appendix D.

Encoder / variant	kHz	Win. (s)	K-class	Filter
AVES-HuBERT DAPT	16	5	89.66	**99.10**
AVES-HuBERT DAPT (32 native, 16 logical)	32/16	2.5	88.43	98.60
WHISPER-tiny (DORI backbone)	16	5	**90.67**	98.30
WAVLM base	16	5	86.01	98.40
WAV2VEC 2.0 base	16	5	86.52	96.40
animal2vec baseline (lr $= 10^{-4}$)	16	5	66.92	95.50
animal2vec Modern2026 ALiBi + clip2	16	5	62.70	92.30
animal2vec whole-listener scratch	32	2.5	60.16	96.60
animal2vec Modern2026 arch + RoPE	16	5	63.50	93.09
animal2vec Modern2026 RoPE + clip2	16	5	61.29	91.59
animal2vec whole-listener refresh	8	10	57.98	91.40
animal2vec multires + dropout	16	5	55.28	92.49
log-mel control (64 mels, mean+std)	16	5	58.60	90.70

Table 4: Sperm-whale click detection (Dominica, file-disjoint holdout, seed 42; 5,766 held-out clicks and 0.721 h of noise). The kHz column gives physical input rate and, where different, the rate the front end assumes (32/16 denotes physical/logical rate); g is the physical embedding interval (ms). SSLUnsupDet is an external unsupervised click detector run at native rate. $q_g = \min(1, 2t/g)$ gives nominal coverage at $t = 2.5$ ms (Sec. 3.4). Encoders are frozen and a 6×64 BiLSTM head is trained. For animal2vec, both collars use stored scans of the same head at each rate, with thresholds 0.81/0.83/0.84 for 8/16/32 kHz originally selected on held-out signal files. FAR/h uses the reported noise-only operating point. Precision is omitted here; $R@P$ is not yet computed. Dashes: not scored at 2.5 ms (two AVES rows) or no fixed embedding grid (SSLUnsupDet); the parenthesised value is the one AVES row scored at 2.5 ms. Textual differences use unrounded scores.

			t=20 ms		t=2.5 ms		
Detector	kHz	g	F1	Rec.	q_g	F1	FAR/h
AVES-HuBERT marine ckpt. A	32/16	10.04	**0.773**	0.816	0.50	—	**0.00**
AVES-HuBERT marine ckpt. B	32/16	10.04	0.760	**0.848**	0.50	(0.150)	**0.00**
animal2vec + BiLSTM	32	1.25	0.722	0.805	1.00	0.659	**0.00**
animal2vec + BiLSTM	16	2.50	0.678	0.753	1.00	0.517	1.39
AVES-HuBERT marine, 16 kHz	16	20.08	0.635	0.648	0.25	—	**0.00**
animal2vec + BiLSTM	8	5.00	0.553	0.574	1.00	0.195	**0.00**
SSLUnsupDet enc. + BiLSTM (untuned)	48	—	0.093	0.082	—	0.016	12.48

5.3 INPUT SAMPLE RATE IS TASK-DEPENDENT

Clicks: F1 rises with rate, and the collar sets the size of the gap. At the common 20 ms tolerance, event F1 rises monotonically $0.553 \to 0.678 \to 0.722$ for animal2vec at 8/16/32 kHz ($+0.168$), and frame-level AUROC moves in step ($0.928 \to 0.984 \to 0.996$). The stored collar scans for the same fitted heads, with thresholds held fixed across collars, give F1 0.195/0.517/0.659 at 2.5 ms. The 32–8 kHz gap is then 0.464, approximately $2.8\times$ larger. Matching tolerance therefore changes the size of the observed advantage; the confounds below mean it does not isolate bandwidth. For AVES, nominal grid coverage at 2.5 ms is 0.50 or 0.25. The one AVES row scored at 2.5 ms (checkpoint B: frame AUROC 0.988, event F1 0.150) points to event extraction and timing rather than frame-level discrimination as the limiting step.

Table 5: Presence on DORI's ONC benchmark (385 files, 159 positive; 5-fold stratified CV, out-of-fold scores). AUROC is DORI's metric; $R@P$ is this benchmark's. Model F1 is at a fixed 0.5 probability threshold. Human rows are the released amateur annotations scored against the same expert reference (positive-class F1); their recall is quoted at their own precision.

Detector	AUROC	F1	$R@P$=0.5	$R@P$=0.8	(P, R)
WHISPER-tiny, max-pool (L3)	**0.870**	0.734	**0.981**	**0.692**	—
WHISPER-tiny, mean-pool (final)	0.844	0.717	0.918	0.654	—
WAVLM base, mean-pool (L2)	0.818	0.714	0.943	0.516	—
WAV2VEC 2.0 base, mean-pool (L3)	0.803	0.684	0.925	0.503	—
amateur annotator 1	—	**0.810**	—	—	(0.99, 0.69)
amateur annotator 2	—	0.762	—	—	—
amateur annotator 3	—	0.618	—	—	(0.51, 0.79)

Interpreting the common-tolerance comparison. The animal2vec rows compare separately trained checkpoints and heads; rate, bandwidth, context and pretraining data vary together. The two AVES rows at 32 kHz native (0.773, 0.760) are two checkpoints of one lineage and both carry a non-random `finetune_from_model` initialisation, so they are *not* a DAPT-vs-scratch ablation. AVES at 16 kHz (0.635) vs. 32 kHz native (0.773) also differs in pretraining manifest (2.55M vs. 4.85M segments) and frame interval, so the +0.138 is a lineage difference, not a resampling ablation; the matched ablation needs no new pretraining and is left to follow-up work. The untuned 48 kHz SSLUnsupDet row is a wide-band reference, not evidence for the bandwidth claim.

Bandwidth intervention at a fixed input rate. With a newly fitted BiLSTM head over the frozen 32 kHz animal2vec encoder, low-pass filtering at 3.5 kHz reduces event F1 from 0.7246 to 0.3999 at a 20 ms collar while the detector, input rate and 1.25 ms embedding grid remain fixed. Because the head was trained on unfiltered audio, filtering also shifts the input distribution, so the loss is an upper bound on the value of the removed band for this detector. Appendix F.1 gives the protocol, full results and paired file-bootstrap intervals.

Call type: a small gain in the band-limited probe. The historical 32 kHz animal2vec run scores below the 16 kHz baseline on K-class (60.16 vs. 66.92), but above it on filtering (96.60 vs. 95.50); those recipes also differ in context and training history. On the native 48 kHz recordings only 2.7% of call energy lies above 8 kHz; a log-mel probe restricted to 0–16 kHz improves on 0–8 kHz by +0.015 macro-F1 (0.554 → 0.569), and the 8–16 kHz band alone reaches 0.260. A band-masking attribution of the frozen probes (Sec. 5.6) places all three tested encoders' reliance in the 0.5–3.5 kHz band. **Bandwidth should be evaluated per task**: the historical 32 kHz model gains 0.044 click F1 over its 16 kHz counterpart, and the fixed detector loses 0.325 when the band above 3.5 kHz is removed, while the call-type bandwidth experiment gains 0.015 macro-F1. A common high-rate front end remains an option, but its value has to be measured per task. The effect of DORI's 1 kHz high-pass on call-type recognition remains untested.

5.4 PRESENCE DETECTION ON THE PRIOR SYSTEM'S OWN TEST SET

The metric and the operating point decide the comparison. Read by F1, the model (0.734) sits below the best annotator (0.810). At $P \geq 0.5$, model recall is 0.981, compared with annotator 3's 0.79 at $P = 0.51$. At $P \geq 0.8$, it reaches 0.692, close to annotator 1's 0.69, although that annotator operates at the stricter $P = 0.99$. A single headline metric hides these trade-offs between missed detections and false alarms. **Human-level is a region, not a line.** The three annotators span P=0.99/R=0.69 to P=0.51/R=0.79 on identical audio. **Pooling is encoder-specific.** Max-pooling helps WHISPER-tiny on every column (+0.026 AUROC, +0.063 $R@P$=0.5) yet hurts WAVLM (0.818 → 0.781) and WAV2VEC 2.0 (0.803 → 0.773), so aggregation must be swept per encoder. The gap between encoders is far wider at the strict floor ($R@P$=0.8: 0.692 vs. 0.503) than at the loose one.

Table 6: Matched-exposure AVES field detection (D044): 17 recordings, 613 reference events. Event scores use a 0.2 s onset collar and offset tolerance $\max(0.2\,\mathrm{s}, 0.5 \times \mathrm{duration})$. Metrics pool outer-held-out recordings per refit, then average three seeds; ± is seed SD. The frozen champion is selected on inner event F1. q denotes positive field presentations per training event.

Adaptation / exposure	Event F1 (%)	P (%)	R (%)
Frozen linear, q=10	40.27 ± 0.37	46.99	35.24
Full FT, q=10	36.27 ± 4.32	49.11	28.77
Frozen linear, q=60	38.08 ± 0.15	44.21	33.44
Tuned linear, q=60	39.92 ± 0.08	44.79	36.00
Residual MLP, q=60	39.20 ± 0.13	47.24	33.50
Frozen champion, q=60	39.74 ± 0.13	45.22	35.45
Frozen + dropout, q=60	38.06 ± 0.42	44.36	33.33
Full FT, q=60	41.18 ± 1.97	54.18	33.22
LoRA, q=60	41.51 ± 0.15	50.54	35.24

5.5 Label provenance: are propagated labels exchangeable?

Rung T8 asks what a reported gain owes to labels a model wrote rather than a person. The DCLDE ecotype corpus marks 6,436 of 15,407 labels (42%) as propagated pseudo-labels. The two kinds of label were selected by different mechanisms, so pooling them needs a check. A classifier given *only* metadata (segment duration and hydrophone coordinates, no audio) predicts human-vs-machine at AUROC 0.612 under a site-grouped split, so provenance is partly recoverable. The skew has a direction: pseudo-labels are 85.5% SRKW against 76.1% for human labels, with fewer Bigg's (13.4% vs. 22.3%), and one contributor supplies 14.7% of propagated labels against 7.2% of human ones. Pooling therefore changes the class and site composition of the training data. What that does to recognition needs the matched retraining arm (human-only vs. human-plus-propagated), which needs per-file audio and is deferred (Appendix G).

5.6 Field generalisation and a bandwidth diagnostic

Adaptation depends on training exposure. Matched arms use identical field/laboratory sample streams, batch 384, and a common warmup through $q5$ followed by constant learning rates; $q10/q60$ are prefixes of trajectories trained to $q80$. At $q10$, frozen-head/full-FT F1 is $40.27/36.27\%$; at $q60$ it is $38.08/41.18\%$. The exposure-by-adaptation interaction is $+7.10$ F1 points (paired recording-bootstrap 95% CI $[3.05, 11.47]$). Selecting the frozen-head exposure by inner event F1 rather than cross-entropy gives 39.72% versus 36.85% F1.

Strong frozen heads narrow the adaptation gap. At $q60$, full FT exceeds the inner-selected frozen champion by 1.43 points (95% CI $[-1.92, 5.09]$); LoRA exceeds full FT by 0.33 points ($[-2.07, 2.71]$). These contrasts establish neither a clear advantage nor equivalence. Earlier D040/D043 recipes and their larger recipe-dependent gains are retained separately in Appendix F.

Calibration distinguishes systems with close F1. At a validation-calibrated precision target of 0.5, full FT supports $17/17$ recordings, with policy recall 28.44% and outer precision 59.90%; LoRA supports $11/17$, with recall 18.92% and precision 60.10%. The paired recall difference is 9.52 points (95% CI $[2.22, 16.77]$). The frozen champion has no feasible inner calibration and abstains on all folds. Policy recall counts all reference events, including those in abstained recordings; it is not oracle recall from the outer score curve. At target precision 0.8, all completed D044 conditions have empty admissible oracle sets and their calibrated policies abstain. Appendix E specifies exposure, selection and calibration; these are retrospective comparisons on a historically studied corpus.

Why the rate effect is task-dependent. A band-masking attribution of the frozen probes (Fig. 2, Appendix H) shows all three tested encoders concentrate reliance in the 0.5–3.5 kHz call-fundamental band, consistent with the direct spectral measurement; the diagnostic covers only the 0–8 kHz band that 16 kHz encoders retain. Call-type probes lean on the low band; click detection, tested separately, loses most of its F1 when the high band is removed.

Whale-specific features on continuous recordings. Google's Multispecies Whale Model (GMWM) recognises vocalisations from eight whale species, including killer whales, but its native

outputs do not identify the 11 Avacha K-call types (Harrell, 2024). We used its released 1,280-dimensional pre-classifier embedding (Google Research, 2024) to train a class-weighted logistic-regression classifier for the K types and noise. We divided the labelled short clips into training, validation and held-out test files, using validation to select the regularisation strength. The classifier reached macro-F1 0.974 on the held-out clips. We then applied it unchanged to the 17 continuous 2016 recordings.

Table 7: GMWM on non-overlapping 5-s windows from the 17 Avacha-Gulf 2016 recordings. Call-presence scores use all windows; K-type scores exclude windows containing multiple K types. Indented rows are the degenerate baselines required by Table 1.

Evaluation	Windows	Score
Call vs. noise, average precision	1,150	0.477
positive-window prevalence (AP baseline)	1,150	0.387
Call vs. noise, balanced accuracy (τ=0.5)	1,150	0.565
chance	1,150	0.500
Noise or one K type, macro-F1	837	0.070
all-noise majority predictor, macro-F1	837	0.076
One K type only, macro-F1	132	0.009

Expert event intervals give 705 windows without a K call, 132 with one K type, and 313 with multiple K types. The last group is included only in the call-presence scores, computed from $1 - P(\text{noise})$; balanced accuracy uses a fixed threshold of 0.5. Among the 132 single-type windows, the classifier assigned the correct K type once and predicted noise 104 times. Its strong short-clip result therefore does not carry over to continuous audio: call-presence average precision is 0.477 against a positive-window prevalence of 0.387, 12-way macro-F1 on the 837 noise-or-single-type windows is 0.070, below the 0.076 of a predictor that labels every window noise, and single-type macro-F1 is 0.009 (Table 7).

On the same 17 recordings, our lab-only AVES detector reaches event F1 0.338, increasing to 0.488 after adaptation using field labels (Table 10). The metrics describe different outputs: AVES detects event intervals, whereas GMWM assigns one label to each 5-s window. Even at this coarser resolution, the clip-trained GMWM classifier rarely identifies the K type. The result shows why success on isolated calls cannot stand in for field performance in the benchmark.

6 Limitations and Conclusion

Recording hardware, deployment conditions and annotation conventions vary across the archives. Controlled downsampling leaves sample rate partly confounded with dataset identity, and the largest high-rate archives lack labels. The historical comparison reports two collars; the 2.8× ratio describes the change in the 32–8 kHz gap between 20 and 2.5 ms. The fixed-detector intervention separately tests input-band removal. Inter-annotator agreement is reported where labels permit, with and without ambiguous clips (Appendix I).

We propose MarineMammalBench, a benchmark and training protocol in which sample rate, resampling strategy, retained bandwidth and embedding grid are logged variables rather than hidden defaults. The first generation reveals complementary strengths: marine animal2vec leads official BEANS Watkins at 92.04% accuracy, validation-selected readout substantially improves Whisper transfer, while AVES-bio leads the linear HICEAS presence comparison. Historical Whisper-tiny and AVES-DAPT results lead call type and filtering, respectively. Matched field adaptation reveals an interaction with training exposure: strong frozen heads narrow the F1 gap, while calibrated coverage distinguishes full fine-tuning from LoRA. Presence detection exposes distinct model–human trade-offs across metrics and precision requirements. GMWM provides a field-transfer counterexample: a whale-specific encoder classifies held-out short clips well, yet its clip-trained head predicts noise for most single-K windows in continuous recordings. The controls also produce findings of their own: input-band restriction to 3.5 kHz reduces click F1 by 0.325 in a fixed detector, whereas changing collar changes the historical 32–8 kHz gap by 2.8×. The annotation-geometry baseline nearly doubles the majority-class floor on the ecotype labels without any audio (0.398 vs. 0.215),

exposing an evaluation shortcut. The result is a working protocol for checking what information reaches a marine SSL encoder, how its outputs are scored, and what the comparisons can support.

Reproducibility statement

The code and configuration behind every number in Sec. 5 will be released with the benchmark: the frozen-probe shim that puts public encoders and the log-mel control through the same probe, split and metric path as the in-domain rows; the presence-rung evaluation including human-annotator scoring; the uncertainty estimator; the annotation-geometry control (about two minutes on a laptop, no GPU, no audio); and the label-provenance audit. Job configurations are released as run. Split definitions, seeds and manifests for every reported table are listed in Appendix D. Trained marine SSL checkpoints, pretraining recipes, the full unlabelled-pool manifests and model cards are deferred to follow-up work; where corpora cannot be redistributed we release manifest builders and access instructions.

Ethics statement

The benchmark uses passive acoustic recordings of wild animals and public archival data; no human subjects are involved beyond published annotator columns released with an existing dataset. The Avacha-Gulf killer-whale corpus that anchors T3 and the field experiment was collected and annotated by a laboratory associated with this work; annotation was not blinded to the hypotheses of this benchmark and no independent inter-annotator statistic exists for it, so results on that corpus should be weighed accordingly. The authors declare no other competing interests.

AI use statement

Large language model tools were used to assist with editing, LaTeX formatting, consistency checking of reported numbers, and drafting of analysis scripts. All experimental design, data handling, results and claims were produced and verified by the authors, who take full responsibility for the content.

References

A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *NeurIPS*, 2020.

C. Bergler, H. Schmitt, A. Maier, S. Haderlein, and E. Nöth. ORCA-SPOT: An automatic killer whale sound detection toolkit using deep learning. *Scientific Reports*, 2019.

C. Bergler, F. Tenorio, Y. Stylianou, and E. Nöth. ANIMAL-SPOT: An open, generalizable deep learning framework for animal sound detection and classification. *Scientific Reports*, 12:21485, 2022.

P. C. Bermant, M. M. Bronstein, R. J. Wood, S. Gero, and D. F. Gruber. Deep machine learning techniques for the detection and classification of sperm whale bioacoustics. *Scientific Reports*, 9:12588, 2019.

Ç. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulović. A framework for the robust evaluation of sound event detection. *ICASSP*, 2020.

S. Chen, C. Wang, Z. Chen, et al. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Topics Signal Process.*, 16(6):1505–1518, 2022.

S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei. BEATs: Audio pre-training with acoustic tokenizers. *ICML*, 2023.

DCLDE 2024 organizing committee. DCLDE 2024 Detection, Classification, Localization, and Density Estimation workshop challenge dataset. https://www.dclde2024.com/data-sets/, 2024.

Earth Species Project. What matters for bioacoustic encoding. *arXiv:2508.11845*, 2025.

O. A. Filatova, T. V. Ivkovich, M. A. Guzeev, A. M. Burdin, and E. Hoyt. Social complexity and cultural transmission of dialects in killer whales. *Behaviour*, 154(2):171–194, 2017.

J. K. B. Ford. A catalogue of underwater calls produced by killer whales (*Orcinus orca*) in British Columbia. *Canadian Data Report of Fisheries and Aquatic Sciences*, 1987.

R. Gibb, E. Brownsword, R. G. Glover-Kapfer, and K. E. Jones. Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. *Methods in Ecology and Evolution*, 10(2):169–185, 2019.

Google Research. Multispecies Whale Model. Kaggle Models, 2024. https://www.kaggle.com/models/google/multispecies-whale.

G. Gubnitsky and R. Diamant. Detecting the presence of sperm whales echolocation clicks in noisy environments. *arXiv:2401.00900*, 2024.

G. Gubnitsky, Y. Mevorach, S. Gero, et al. Automatic detection and annotation of sperm whale codas. *arXiv:2407.17119*, 2024.

S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. *ACL*, 2020.

M. Hagiwara. AVES: Animal vocalization encoder based on self-supervision. *ICASSP*, 2023.

M. Hagiwara et al. BEANS: The benchmark of animal sounds. *ICASSP*, 2023.

J. Hamer et al. BIRB: A generalization benchmark for information retrieval in bioacoustics. *arXiv:2312.07439*, 2023.

L. Harrell. Whistles, songs, boings, and biotwangs: Recognizing whale vocalizations with AI. *Google Research Blog*, 2024. Online article.

P.-Y. Huang, H. Xu, J. Li, A. Baevski, M. Auli, W.-N. Hsu, D. Harwath, and J. Glass. Masked autoencoders that listen. *NeurIPS*, 2022.

O. Mac Aodha et al. Bat detective: Deep learning tools for bat acoustic signal detection. *PLOS Computational Biology*, 2018.

A. Mesaros, T. Heittola, and T. Virtanen. Metrics for polyphonic sound event detection. *Applied Sciences*, 6(6):162, 2016.

B. Nestor, B. Yao, J. Moore, J. Kanes, et al. Positive-unlabelled active learning to curate a dataset for orca resident interpretation. *arXiv:2602.09295*, 2026.

A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. *ICML*, 2023.

L. Rauch et al. BirdSet: A multi-task benchmark for classification in avian bioacoustics. *NeurIPS Datasets and Benchmarks Track*, 2024.

D. Robinson, M. Miron, M. Hagiwara, and O. Pietquin. NatureLM-audio: An audio-language foundation model for bioacoustics. *arXiv:2411.07186*, 2024.

J. C. Schäfer-Zimmermann, V. Demartsev, B. Averly, et al. animal2vec and MeerKAT: A self-supervised transformer for rare-event raw audio input and a large-scale reference dataset for bioacoustics. *Methods in Ecology and Evolution*, 2024.

D. Stowell. Computational bioacoustics with deep learning: A review and roadmap. *PeerJ*, 10:e13152, 2022.

J. Turian et al. HEAR: Holistic evaluation of audio representations. *NeurIPS 2021 Competitions and Demonstrations Track*, 2022.

B. van Merriënboer, V. Dumoulin, J. Hamer, L. Harrell, A. Burns, and T. Denton. Perch 2.0: The bittern lesson for bioacoustics. *arXiv:2508.04665*, 2025.

APPENDIX

A WORKED EXAMPLE OF THE FOUR LOGGED QUANTITIES

Consider one Watkins clip recorded at 96 kHz containing a sperm-whale click train with energy to 30 kHz. Three pipelines applied to the same clip retain three different bands:

Pipeline	Orig. SR	Model SR	Resampling strategy	Retained band
wav2vec 2.0 default	96 kHz	16 kHz	polyphase downsample, anti-alias LP @ 8 kHz	0–8 kHz
Perch 2.0 default	96 kHz	32 kHz	polyphase downsample, anti-alias LP @ 16 kHz	0–16 kHz
Log-mel ResNet-18 high-SR control	96 kHz	96 kHz	none (native)	0–48 kHz

The first discards the entire 8–30 kHz band before the encoder sees the signal; the second discards 16–30 kHz; the third retains all of it. Sample rate alone does not capture this: the resampling strategy and the model-imposed input rate are independent levers, so all four quantities are logged.

B DATASETS AND ENCODERS

Table 8: Benchmark datasets and roles. Hours and rates are from the project data manifest; per-dataset manifests fix exact values.

Dataset	Duration	Sample rate	Role	Label type
Labelled, included in v1 (access confirmed):				
Manual sound-filtering set (Watkins master tapes + Orcasound)	∼15 h	mixed	sound vs. noise/artifact triage (T1, specified; catalogue proxy reported)	binary clip/segment
Watkins MMSDB	∼5 h	up to 166 kHz	species classification (T2)	multiclass clip
Watkins sperm-whale subset	—	up to 166 kHz	high-frequency click stress test	clip / event
Orcasound labelled	∼1,900 h	44.1 kHz mostly	orca presence/call/species; SSL pretraining	binary or multiclass
Avacha-Gulf killer-whale corpus (Filatova et al., 2017)	11 call types + noise, ∼45k clips; 17 field recordings (1.2 h)	48 kHz	call type (T3); field detection	clip-level call type; event-level field annotations
Voices in the Sea	0.47 h	5–48 kHz	small semantic validation set	species / sound type
DCLDE 2022/2024	—	44.1–48 kHz	orca detection; DORI anchor; geometry control	binary or multiclass
Dominica sperm-whale click set	82 files (39 signal, 43 noise)	96+ kHz	click detection (T4, Table 4)	event-level times of arrival
BEANS HICEAS	677 parent chunks; 7,427 windows	22.05 kHz	minke-whale window presence (T5)	binary targets from event intervals
DORI ONC presence benchmark	385 × 5 min	—	presence (T5)	expert + 3 amateur columns
Unlabelled, SSL pretraining pool:				
Orcasound archive	∼14,000 h	48 kHz	in-domain SSL pretraining	—
NOAA / NCEI PAD	∼773,000 h	2–500 kHz	in-domain SSL pretraining; spectral stress tests	—
Pacific Sound (MBARI)	∼85,825 h	256 kHz	high-rate SSL pretraining; bandwidth controls	—
ONC hydrophone archive	∼1,300,000 h	4–256 kHz	in-domain SSL pretraining; cross-deployment	—
Planned, conditional on access:				
DCLDE-2027 evaluation set	—	mixed / high-SR	held-out challenge; head-to-head with DORI	challenge labels
CETI extended sperm-whale data	—	96+ kHz	extended high-rate click track	event or presence

Table 9: Encoder baselines and SSL training variants. Public checkpoints use their release default input rate; for in-domain variants the input rate is an experimental variable.

Model	Domain	Representation	Default rate
WAV2VEC 2.0	speech	SSL encoder	16 kHz
WAVLM	speech	SSL encoder	16 kHz
BEATS	general audio	SSL / acoustic tokenizer	16 kHz
AUDIOMAE	general audio	masked autoencoding	16 kHz
AVES	animal vocalisation	SSL encoder (HuBERT)	16 kHz
PERCH 2.0	bioacoustics	foundation embedding model	32 kHz
ANIMAL2VEC	animal vocalisation	SSL encoder (data2vec)	8 kHz or checkpoint-dependent
WHISPER-tiny	speech / general audio	supervised encoder (frozen)	16 kHz
Log-mel RESNET-18 (16 kHz / high-SR)	task-specific	supervised CNN reference / positive control (planned)	16 kHz / native
Marine SSL variant A	marine PAM	SSL encoder on unfiltered archives	8/16/32/high-SR
Marine SSL variant B	filtered marine PAM	SSL encoder retrained after filtering	8/16/32/high-SR

The first in-domain pretraining corpora drawn from the pool are an 8 kHz / 10 s corpus of 3,089.8 h (NOAA ONMS 1,784.8, Orcasound 576.7, Pacific Sound 432.0, ONC 296.3) and a 16 kHz / 5 s corpus of 3,542.3 h (NOAA 1,784.6, Orcasound 1,047.5, Pacific Sound 425.1, ONC 285.1), the latter differing chiefly in a fully recovered Orcasound archive. Source files are resampled once at download and the originals are not retained, so higher-rate corpora are rebuilt from source rather than upsampled. ANIMAL2VEC's 8 kHz default removes essentially all cetacean ultrasonic content and is reported as a deliberate low-bandwidth animal-SSL control. A supervised log-mel RESNET-18 in 16 kHz and native-rate variants is planned as the trained non-SSL reference.

C TASK SPECIFICATION, T6–T9

T6 holds out whole *encounters*, not deployments, because simultaneous detections of the same animals at nearby hydrophones otherwise leak across the split. T7 stratifies by range using the corpus range annotations, because high-frequency absorption scales roughly with f^2: retained bandwidth can only pay off at short range, and pooling near- and far-field encounters averages a real effect against a physical null. T8 retrains with model-propagated labels removed and evaluates on held-out human labels. T9 fixes recall and compares joules and latency, so an efficiency claim cannot be won at an arbitrary operating point.

D EXPERIMENTAL PROTOCOL DETAILS

Splits. DCLDE: official challenge splits when provided. Orcasound and orca call-type data: by recording session and year, with a temporal split (train before 2021, validation 2021, test 2022) where possible, following Nestor et al. (2026). Watkins: exact official BEANS CSV membership over 31 classes, with 1,017/339/339 train/validation/test rows; the released split already excludes Weddell seal. We introduce no further exclusions or reshuffle. The release contains 25 exact-byte pairs with conflicting labels, including 13 cross-split pairs; 279/339 test rows map to the historically available confit train pool. This public test is not a fresh blind or verified source-disjoint holdout; pretraining overlap is unknown. Avacha-Gulf call type and filtering: clips are 48 kHz catalogue cuts of single calls (11 call types) plus an ambient-noise class from the same recordings; the group key for 5-fold GroupKFold is the recording date–time token parsed from the file name. The K-class rows in Table 3 were produced with a fixed class list of the 11 call types *plus* the noise class (12-way); a later revision of the harness drops noise from the call-type probe (11-way), and the class list of each run is now logged with its result. The filtering set draws up to 700 clips per class (seed 42),

pools all call types as *signal*, and truncates signal and noise to 500 each (seed 0); the resulting 1,000 file identifiers and group keys are released as a manifest. Noise here is the corpus's own ambient class, a proxy for the dedicated ∼15 h manual filtering set, which is specified but not yet used. The grouping has two known mismatches: the file-name key is not a verified recording identifier (175 of the 1,000 manifest entries carry a non-standard token, and one noise recording appears under two names), and the harness prefixes noise groups, so a recording contributing both calls and noise is treated as two groups (2 such recordings, 5 clips, in the manifest). The filtering fold assignment is therefore approximately, not strictly, recording-disjoint; a canonical recording map from the source archive will replace the file-name key in follow-up work. A balanced 500/500 probe also does not measure false-alarm rate on long field recordings, which the field experiment (Sec. 5.6) addresses separately. Sperm-whale clicks: file-disjoint split with seed 42 (31/8 signal and 34/9 noise files); no split is made by individual click. DORI ONC presence: 5-fold stratified cross-validation over 385 files with out-of-fold scores. AVES field experiment: outer recording folds of 6/6/5 recordings, with repeated inner 3-fold CV for model and threshold selection. For GMWM, the labelled short clips were split by file; the 17 continuous field recordings were used only for evaluation (Sec. 5.6). Per-dataset split files are released with the manifest builders. **Probes.** Linear (logistic regression or one linear layer) and shallow MLP (one hidden layer, dropout), hyperparameters selected on validation and fixed for test; for detection, a lightweight temporal head (6×64 BiLSTM) is trained on frozen features and reported separately. **Metrics.** Official BEANS Watkins: accuracy primary and macro-F1 secondary over all 31 classes. BEANS HICEAS uses window rank AP; other classification tasks retain macro-F1 primary. Detection: $R@P{=}0.5$, $R@P{=}0.8$ primary; AUROC, AP, event F1 with collar sweep, segment F1 and PSDS secondary. Presence-rung model F1 is computed at a fixed 0.5 out-of-fold probability threshold; human positive-class F1 is $2PR/(P+R)$ from the released annotation columns. **Official BEANS frozen probes.** The pinned BEANS release is `b278b1d96b3fb7ad4d14eb7453fc61ab071bc897`. For Watkins, mono waveforms are resampled to 44.1 kHz before taking the first up to 3 s (or 5 s in the sensitivity arm), then converted to each encoder's input rate with sinc resampling; short clips are right-padded. The AVES row uses original AVES-bio without marine DAPT. AVES-bio and animal2vec16 use 16 kHz/5 s model context, Whisper-tiny 16 kHz/30 s, and Perch 2.0 32 kHz/5 s. AVES readouts are blocks 1/6/12; Whisper uses blocks 1/2 and its post-final LayerNorm; animal2vec uses actual blocks 2/12 and native final output. Perch uses its native embedding. Animal2vec loads the delivered BEST checkpoint at 129,000 updates and explicitly disables masking during inference. Pooling follows valid frames where supported; native normalization and context remain model-specific. Log-mel uses 64 band means plus standard deviations; the random control is a seed-17 frozen three-layer CNN. A train-only StandardScaler precedes balanced logistic regression with $C \in \{0.1, 1, 10\}$, LBFGS, tolerance 10^{-4} and at most 5,000 iterations. On Watkins, validation accuracy selects readout/C; ties prefer the declared readout order, then smaller C. The native-final policy selects C within that readout. Three-readout encoders have nine candidates; single-readout controls and Perch have three. Selections and probes are saved before test extraction, without train-plus-validation refitting. Code, model hashes and locked dependencies accompany the run artifacts. **HICEAS windows and heads.** The official release has 407/135/135 train/validation/test parent chunks at 22.05 kHz and 4,457/1,485/1,485 windows. Splits precede 10 s windows with 5 s hop; a short training parent remains listed but yields no eligible window. Targets follow official overlap > 0.2 relative to the smaller event/window duration, including full-window containment. Resampling retains parent context at crop boundaries. Each window averages two consecutive 5 s embeddings per readout. Train-fitted scaling and balanced logistic regression use the same C grid as Watkins, with readout/C selected on validation rank AP. AP averages precision at each ranked positive using float32 scores in fixed row order and descending Torch CPU nonstable sort, matching pinned BEANS code; it is window AP, not event AP or an eleven-point interpolated metric. Linear probes rank continuous decision logits; the MLP ranks positive minus negative logit. The secondary MLP reuses the linear validation-selected readout and train-only scaler: 256 hidden units, ReLU, dropout 0.1, weighted cross-entropy, AdamW with learning rate 10^{-3} and weight decay 10^{-4}, batch 128, 200 epochs, seeds 17/29/43. Validation AP selects each seed's checkpoint; all seeds freeze before their test scoring. The same fixed MLP recipe was applied to all six feature families; the reported Perch result is the mean and sample SD, not a confidence interval. Exact-file duplicates are absent in HICEAS, but partial recording overlap and encounter independence remain unverified; neither dataset is claimed pretraining-disjoint. **AVES field adaptation.** The main comparison uses original AVES-base-bio at mono 16 kHz, 1.0 s windows and 0.1 s hop, per-window mean/std normalization and last-layer temporal-mean features. A weighted laboratory-trained linear probe initializes the shared head. Matched-exposure optimiza-

tion, nested selection, stronger frozen controls, LoRA and calibration are specified in Appendix E; the earlier recipes are in Appendix F. **In-domain SSL pretraining.** The animal2vec-style encoders are data2vec-multi "large" models (depth 16, embedding 1,024, 16 heads, $\approx$ 315M parameters) trained with the fairseq 0.12.2 recipe on one node with 2$\times$ NVIDIA A100-SXM4-80GB GPUs (distributed data parallel, bf16 mixed precision). Per GPU, 320,000 tokens per batch with 10 gradient-accumulation steps and 3 clones per clip give an effective batch of about 80 clips per optimizer update. Adam (β_1=0.9, β_2=0.98, ϵ=10^{-6}, weight decay 0.01), peak learning rate 10^{-4} with 10,000 linear warm-up updates and cosine decay over a nominal 120,000-update schedule, gradient-norm clipping at 1.0, seed 1. The 8 kHz/10 s run (torch 1.13.1, CUDA 11.7) was resumed at update 13,581 from an earlier learning-rate-5×10^{-5} run and stopped at update 54,324 (five epochs, 184.7 h wall time at $\approx$ 12.2 s per update); the 16 kHz/5 s run (torch 2.2.2, CUDA 12.1) ran to update 129,000 ($\approx$ 140 h at $\approx$ 3.9 s per update). Checkpoints were saved every 1,000 updates; historical checkpoints were retained by the original validation procedure. The new BEANS comparison fixes the delivered 129,000-update BEST checkpoint. No pretraining seed was repeated. **Compute.** SSL pretraining, one-time embedding extraction (cached) and probe training are separated so total cost is dominated by SSL: about 370 and 280 A100 GPU-hours for the two runs above.

E MATCHED AVES FIELD PROTOCOL

Data and initialization. The accepted Avacha material contains 45,429 laboratory files, comprising 33,201 call clips across 11 K-types and 12,228 noise clips; 17 field recordings contain 613 expert events over 1.2213 h. The classifier has 12 outputs, including noise; the event metric uses the 11 call types. The laboratory probe supplies a shared warm start and fixed feature scaler. Class, source-stratum, and content-family sampling weights preserve the supplied laboratory pool and reduce repeated-copy dominance. Positive training windows are centered on annotations; background windows exclude reference events. Outer inference is dense over full recordings.

Splits and refits. Outer recording folds are fixed at seed 0 with 6/6/5 recordings and 187/275/151 reference events. Inner three-fold partitions are repeated with split seeds 1/2. Each outer selection aggregates the two complete inner out-of-fold predictions. The main tuning seed is 0; final refit seeds are 0/1/2. Final trajectories run to $q80$ and provide exact prefix checkpoints to fixed-exposure and selected-exposure evaluations. The frozen champion aliases the family selected within that outer fold. Exposure selection compares $q \in \{5, 10, 20, 40, 60, 80\}$ by inner cross-entropy or inner event F1; outer scores do not choose the checkpoint.

Exposure. For each training fold, q is positive field presentations divided by the number of training events. Each update samples 96 positive field, 96 background field and 192 laboratory windows, so $U(q) = \lceil qN_{\text{train events}}/96 \rceil$. Endpoints $q10/q60$ are prefixes of $q80$ trajectories. Frozen encoders use evaluation mode, except the explicit dropout control; full FT and LoRA retain ordinary encoder dropout. SSL masking is disabled. **Optimization.** Batch 384 contains equal numbers of positive and background field windows and 50% laboratory windows. Accumulation is 1. Field and laboratory CE each receive weight 0.5, with label smoothing 0.05. AdamW uses weight decay 10^{-4} excluding bias/norm parameters, global gradient clipping 1.0, BF16 autocast, and FP32 parameters and states. Head LR is 10^{-4}; full-FT backbone LR is 10^{-5}. At zero-based update u, the common warmup multiplier is $0.1 + 0.9\min((u+1)/U(5), 1)$. The scheduler then remains constant. Sampler, model, and diagnostic random streams are separated. At a given fold/exposure, matched arms have identical update counts and sample traces. Ordinary dropout and global gradient clipping are part of each adaptation recipe.

Tuned linear and residual heads search $\{3\times10^{-5}, 10^{-4}, 3\times10^{-4}, 10^{-3}\}$ at $q60$. The residual head adds a 384-unit GELU branch to the laboratory linear head; its output layer starts at zero. The frozen champion searches the union of both families on inner event F1. LoRA inserts rank 8 updates with alpha 16 into query/value projections in all 12 transformer layers, adding 294,912 adapter weights. It trains the linear head and adapters with the base encoder weights frozen. Adapter LR candidates are $\{10^{-5}, 10^{-4}, 10^{-3}\}$. Selection is at $q60$ on inner event F1. All original encoder parameters, including the convolutional frontend, are trainable in full FT.

Threshold and event contract. Detection candidates are $\{0.01, 0.02, \ldots, 0.99, 0.995, 0.999\}$, edge candidates are $\{0.05, 0.10, 0.20, 0.30, 0.40, 0.50\}$, and edge must be smaller than detection. The 451 pairs are ordered detection-major, edge-minor for ties. Event conversion uses a length-5 median

Table 10: Earlier batch-32 AVES field recipe (D040): detection on 17 Avacha-Gulf 2016 recordings, 613 expert events; class-aware `sed_eval` event F1 with a 0.2 s onset collar and offset tolerance $\max(0.2\,\text{s}, 0.5 \times \text{duration})$. Nested recording-level CV selects adapted models and thresholds internally. Adapted rows average three seeds; SD measures seed variation and 95% CIs use 2,000 recording-bootstrap draws. Lab-only weights use no field labels, but their thresholds are calibrated on labelled outer-training recordings.

Detector	F1 (95% CI)	P	R	Seed SD
Lab-only AVES + linear probe	0.338 [0.286, 0.390]	0.284	0.416	—
AVES, head-only adaptation	0.344 [0.290, 0.404]	0.300	0.402	0.008
AVES, full fine-tuning	**0.488** [0.403, 0.575]	0.600	0.411	0.020

filter, strict threshold comparisons, interpolated boundaries, union of overlapping/touching regions of the same class, and minimum duration 0.15 s. The matcher uses maximum-cardinality one-to-one assignment; collar comparisons include a numerical tolerance of 10^{-12} s. Class-agnostic evaluation relabels the same event lists while retaining their boundaries and multiplicity. The collar diagnostic varies onset tolerance over 0.1/0.2/0.3 s and retains the offset fraction 0.5.

For each precision floor, calibration requires nonempty predictions meeting that floor in both inner repeats. Among feasible pairs it maximizes mean recall; ties use the declared pair order. The policy emits nothing on infeasible folds. Strict calibrated recall is defined when all three folds are feasible; policy recall always retains all reference events. The oracle grid evaluates one common pair across all outer recordings. An empty admissible set is recorded as unattainable. The LoRA refit with seed 2 has an oracle outcome at $p = .5$ that is unattainable even though its fold-specific F1 operating point has a different precision, because the oracle requires a single shared pair.

Bootstrap. All reported field intervals use 2,000 paired samples of the 17 recording IDs with replacement, RNG seed 0. Repeated IDs contribute their counts multiple times. Pooled metrics are recomputed per seed and averaged over three refits. Calibration thresholds and model weights are fixed in these draws. No infeasible oracle seeds or draws are averaged away.

Completed fresh-probe follow-up. The separate D046 recovery refits identical fresh linear probes on frozen original and adapted encoders. Mean event F1 is 48.71% for E0 and 47.54% for F60; the paired difference is -1.16 points (95% CI $[-5.96, 3.14]$, 10,000 recording resamples). Under the preregistered ± 2-point practical margin this is inconclusive, not evidence of persistent representation improvement or equivalence. These fresh-probe scores use a different procedure from Table 6.

F Earlier AVES field recipes

The D040 and D043 results below retain their original optimization and threshold grids. Their exposure and selection procedures differ from the matched D044 comparison, so score differences across these recipes do not isolate an encoder-adaptation effect.

Adaptation gain under the earlier recipe. On the 17 recordings (Table 10), full fine-tuning raises event F1 from 0.338 to 0.488: +15.01 percentage points (95% paired recording-bootstrap CI $[10.02, 20.83]$). It also exceeds head-only adaptation by 14.41 points $[9.73, 19.69]$; head-only adaptation gives no clear gain over the laboratory model. The improvement comes primarily from precision ($0.284 \to 0.600$), while recall remains similar ($0.416 \to 0.411$). The D043 recipe, with effective batch 384 and one tuning seed rather than batch 32 and repeated tuning, yields head-only/full-FT F1 of $0.348/0.412$. The size of the adaptation gain therefore depends on the optimization and selection recipe; these runs do not isolate batch size.

AVES field adaptation. Table 10 uses the torchaudio AVES-base-bio checkpoint with mono 16 kHz input, 1.0 s windows, 0.1 s hop, per-window mean/std normalization and last-layer mean pooling. A weighted StandardScaler and multinomial L2 logistic probe ($C = 1$) use all 45,429 laboratory clips, with equal class mass and nested weighting by augmentation stratum and content family. This deterministic 12-way probe initializes both adapted arms; noise is excluded from the 11-class event metric. Head-only adaptation freezes the encoder; full fine-tuning updates the entire encoder, in-

cluding its convolutional frontend, and the head. Each effective batch of 32 comprises 8 positive field, 8 background and 16 laboratory windows (microbatch 8, accumulation 4). AdamW uses backbone/head learning rates $10^{-5}/10^{-4}$, weight decay 10^{-4} except bias/norm, label smoothing 0.05 and gradient clipping 1. Validation-CE plateau scheduling and early stopping select inner checkpoints. For each adapted mode, outer fold and seed $0/1/2$, inner 3-fold CV is repeated with split seeds $1/2$. Its six best checkpoints yield two sets of out-of-fold scores; pooled event F1 selects detect/edge thresholds from $\{0.5, 0.6, 0.7, 0.8, 0.9\}$ and $\{0.2, 0.3, 0.4, 0.5\}$, with edge below detect. Final refits use the median best epoch count and transferred inner LR trajectories; outer labels never select parameters or stopping time. Event conversion uses median-5 smoothing, interpolated edges, union of touching same-class regions, minimum duration 0.15 s and the actual 1.0 s window. Class-aware `sed_eval` 0.2.1 uses onset collar 0.2 s and offset tolerance $\max(0.2\,\text{s}, 0.5 \times \text{duration})$. Metrics pool all 17 outer-held-out recordings per seed before averaging; SD is across seeds. Paired percentile CIs use 2,000 recording resamples (seed 0), conditional on the fitted models and this historically studied corpus; recording folds do not establish independent animal encounters. The second recipe uses batch 384 with accumulation 1, one inner tuning seed and three final refit seeds, reducing field fits from 126 to 54. **In-domain SSL pretraining.** The animal2vec-style encoders are data2vec-multi "large" models (depth 16, embedding 1,024, 16 heads, $\approx$ 315M parameters) trained with the fairseq 0.12.2 recipe on one node with 2$\times$ NVIDIA A100-SXM4-80GB GPUs (distributed data parallel, bf16 mixed precision). Per GPU, 320,000 tokens per batch with 10 gradient-accumulation steps and 3 clones per clip give an effective batch of about 80 clips per optimizer update. Adam (β_1=0.9, β_2=0.98, ϵ=10^{-6}, weight decay 0.01), peak learning rate 10^{-4} with 10,000 linear warm-up updates and cosine decay over a nominal 120,000-update schedule, gradient-norm clipping at 1.0, seed 1. The 8 kHz/10 s run (torch 1.13.1, CUDA 11.7) was resumed at update 13,581 from an earlier learning-rate-5×10^{-5} run and stopped at update 54,324 (five epochs, 184.7 h wall time at $\approx$ 12.2 s per update); the 16 kHz/5 s run (torch 2.2.2, CUDA 12.1) ran to update 129,000 ($\approx$ 140 h at $\approx$ 3.9 s per update). Checkpoints were saved every 1,000 updates; the composite-best checkpoint under the frozen probe was retained (Sec. 5). No pretraining seed was repeated. **Compute.** SSL pretraining, one-time embedding extraction (cached) and probe training are separated so total cost is dominated by SSL: about 370 and 280 A100 GPU-hours for the two runs above.

F.1 Fixed-detector bandwidth control

Detector and split. We use the historical 32 kHz animal2vec checkpoint `checkpoint_2_107000.pt`, freeze the encoder and fit one new 6-layer, 64-hidden-unit BiLSTM on its final-layer features for 20 epochs (seed 42). The file-disjoint split is unchanged: 31/8 signal and 34/9 noise files for training/testing, with 5,766 held-out clicks. Training uses the unfiltered input condition, AdamW (learning rate 10^{-3}, weight decay 10^{-4}), batch size 16, and focal loss ($\alpha = 0.8$, $\gamma = 1.5$; automatic positive weighting capped at 100). The same final head is frozen for every evaluation condition; this is a new fit, distinct from the historical heads in Table 4. Encoder tensor hashes before and after head training are identical.

Input intervention. We compare unfiltered audio with nominal low-pass cutoffs of 14, 7.5 and 3.5 kHz. Filtering uses odd-length, linear-phase Kaiser FIR filters at the native recording rate, with a 500 Hz transition and an 80 dB design attenuation target. The nominal cutoff is the transition midpoint; measured responses and coefficients are retained. Native-rate context around each window compensates the filter delay, with zero padding only outside physical recording boundaries. All conditions then use the same resampling to 32 kHz, 2.5 s windows and hops, padding, normalisation, 1.25 ms embedding grid and event extraction. The historical threshold 0.84 is fixed before this fit and is not retuned by condition or collar; its original selection used held-out Signal, so this experiment does not provide a newly calibrated independent test operating point.

Paired uncertainty. We resample the eight signal files jointly across conditions (10,000 bootstrap draws, seed 42), recompute F1 from pooled TP/FP/FN counts, and report percentile intervals for paired differences (Table 12). These intervals condition on one fitted head, assume independent files, and are not adjusted for multiple comparisons. They do not include training-seed variability.

Timestamp and noise diagnostics. The complete analysis evaluates collars of 0.5, 1, 1.25, 2.5, 5, 10, 20 and 40 ms. Rounding predicted timestamps to a common 5 ms lattice, while retaining coincident detections as separate events and leaving ground truth unchanged, changes unfiltered F1 from 0.6615 to 0.6326 at 2.5 ms and leaves its 20 ms F1 unchanged at 0.7246. This intervenes

Table 11: Fixed-detector bandwidth control. Event F1 pools the eight held-out signal files (5,766 clicks); FAR/h uses nine noise files (0.721420 h). All rows share the fitted detector and threshold 0.84.

Input condition	F1 (2.5 ms)	F1 (20 ms)	FAR/h
Unfiltered	0.6615	0.7246	0.00
Low-pass, 14 kHz	0.6624	0.7258	0.00
Low-pass, 7.5 kHz	0.5721	0.6804	0.00
Low-pass, 3.5 kHz	0.3124	0.3999	338.22

Table 12: Filtered-minus-unfiltered F1 differences and paired 95% file-bootstrap intervals. Negative values indicate a loss after filtering.

Low-pass cutoff	Collar (ms)	ΔF1	95% interval
14 kHz	2.5	$+0.0009$	$[-0.0045, +0.0054]$
7.5 kHz	2.5	-0.0894	$[-0.1579, -0.0440]$
3.5 kHz	2.5	-0.3490	$[-0.5410, -0.2486]$
14 kHz	20.0	$+0.0012$	$[-0.0039, +0.0050]$
7.5 kHz	20.0	-0.0443	$[-0.1282, -0.0073]$
3.5 kHz	20.0	-0.3247	$[-0.5735, -0.2299]$

on reported timestamps, not the internal encoder grid. The 3.5 kHz condition produces 244 noise events; the other conditions produce none over the observed 0.721420 h, which bounds their rate below about 4 events/h at the 95% level (rule of three).

Scope. The results show sensitivity to input-band removal for this fitted detector at a fixed input rate and embedding grid. Filtering may also change the input distribution, so the intervention neither decomposes the historical 8–32 kHz gap nor establishes the performance achievable by a detector trained on narrow-band audio. The 14 kHz intervals include zero.

G Label provenance (T8)

The DCLDE ecotype corpus marks 6,436 of 15,407 labels (42%) as propagated pseudo-labels; its release notes state that bulk annotations come from a non-expert (Jaccard ≈ 0.7 against an expert) and that species and ecotype labels were never expert-compared. We examine whether label provenance is associated with observable metadata, as a diagnostic of selection differences. A classifier given only segment duration and hydrophone coordinates predicts human-vs-machine at AUROC 0.612 under a site-grouped split, so provenance is partly recoverable from metadata. Pseudo-labels are 85.5% SRKW against 76.1% for human labels, with fewer Bigg's (13.4% vs. 22.3%); one contributor supplies 14.7% of propagated labels against 7.2% of human ones. The matched retraining arm (human-only vs. human-plus-propagated) needs per-file audio and is not run.

H Band-reliance diagnostic

The attribution is KernelSHAP over 16 frequency bands (200 coalition samples, 10 clips or events per call type): each coalition masks a set of bands in the STFT domain before the encoder, so the same perturbation is applied to every encoder and the scores are comparable across them. It covers the 0–8 kHz band retained at 16 kHz. A full attribution study (contrastive discriminative maps, probe geometry, faithfulness deletion tests) is developed in separate work.

I Extended limitations

Dataset heterogeneity. Archives differ in hardware, deployment context, rate, species coverage and annotation convention; results are evidence about specific tasks and datasets. **Frozen-probe scope.** Full fine-tuning may improve task performance when enough labels exist; the roadmap question is

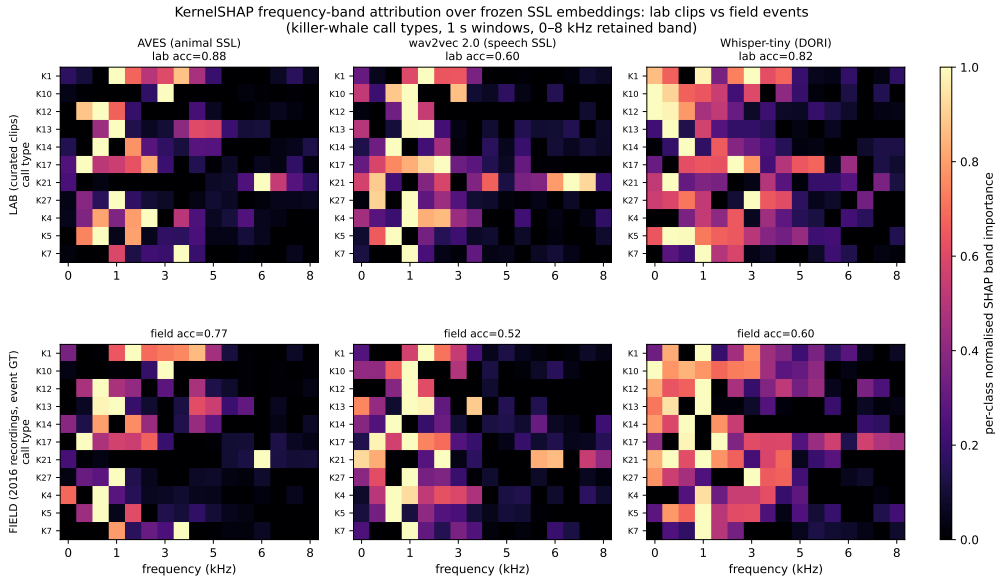


Figure 2: Band-reliance diagnostic: KernelSHAP frequency-band attribution (coalitions mask bands in the STFT domain before the encoder, so comparable across encoders) of a linear probe over frozen embeddings, per-class normalised, for AVES, WAV2VEC 2.0 and WHISPER-tiny on curated call-type clips (top) and field events (bottom); panel titles give probe accuracy in each regime. All three concentrate reliance in the 0.5–3.5 kHz call band.

whether in-domain SSL yields a reusable encoder. **Rate and high-rate-data confounding.** Controlled downsampling and band-limited controls mitigate but do not eliminate residual confounding; the largest high-rate archives are weakly or unlabelled. **Tolerance ladder.** The 20 ms collar is a common reporting choice. Applications requiring millisecond timing should also be evaluated at their required collar with each system's timestamp extraction rule stated; the embedding interval alone does not determine attainable F1 or justify excluding coarse-grid encoders. **Inter-annotator agreement.** Where several annotations exist we report Cohen's κ, Krippendorff's α and event-level boundary agreement, with and without ambiguous clips, as an approximate upper bound on automated agreement. **Open baselines.** A trained supervised CNN and a tuned classical DSP click/coda detector (Gubnitsky and Diamant, 2024; Gubnitsky et al., 2024) are the controls that would show whether the SSL detector's 0.773 is an advance or a number a well-tuned non-SSL detector already reaches.