EarthChirp: a global reference library for insect acoustic recognition and discovery across the audible and ultrasonic spectrum

Marko Melnick[1] (ORCID: 0000-0001-8616-3019)

1. FoundFold, Boulder, CO, USA

Corresponding author: marko.melnick@colorado.edu

Running title: Insect acoustic recognition and discovery

Keywords: passive acoustic monitoring, insects, Orthoptera, Cicadidae, foundation models, few-shot learning, open-set recognition, morphospecies, time-expansion, ultrasonic, Chiroptera, cross-taxon transfer

---

Abstract

1. Passive acoustic monitoring (PAM) is scaling rapidly, but automated recognition for insects lags far behind birds. The dominant recogniser (BirdNET) classifies only ~35 insect species and building bespoke insect classifiers requires labelled training data that does not exist for most taxa.

2. We present EarthChirp, a training-free recogniser for singing insects (Orthoptera and Cicadidae) built on a frozen bioacoustic foundation model (Perch 2.0). EarthChirp pairs a global reference library of 2,012 species (assembled from GBIF, Xeno-canto, SINA and InsectSingers) with a single confidence-gated recogniser that returns a named species where the match is strong and an unnamed morphospecies placed phylogenetically against the reference library (reliably only for the dominant group) where it is not.

3. We find detection and discovery generalise zero-shot to independent data, and show EarthChirp separates insects from vertebrate biophony (AUROC 95%) and recovers morphospecies on an independent 104-species European corpus. Fine species naming does not transfer zero-shot across recording conditions (≈26% open-set), but saturates near 58% with ten local reference clips per species (recording-disjoint). The residual false-positive risk is anthropogenic mechanical noise, not other animals. Using an outgroup prototype we are able to lift insect-versus-noise discrimination from 74% to 93%.

4. EarthChirp is released as a ~29 MB library and classifier head usable with the public Perch model and requires no GPU and no retraining. It detects and discovers insects in any soundscape and adapts to local faunas with a handful of reference clips, which is a deployment recipe rather than a per-site retraining cost. The same frozen-embedding approach reaches beyond the model's 16 kHz ceiling by time-expansion, recovering ultrasonic insects and bats. On the limited open bat audio currently available, we show EarthChirp generalizes to this unrelated ultrasonic taxon (bats) and provisionally places species absent from its reference library. We map the method's capabilities and limits to guide its use in conservation monitoring.

---

1. Introduction

Soundscapes (the combined biotic, geophysical and anthropogenic sound of a place) carry continuous information about ecological state and passive acoustic monitoring (PAM) has matured into a scalable, low-cost way to capture it (Pijanowski et al. 2011; Sugai et al. 2019; Gibb et al. 2019). Singing insects are unusually well suited to acoustic survey, as crickets, katydids, grasshoppers and cicadas advertise more or less continuously and with species-specific songs. Many of these songs reach from the audible range into the ultrasonic, so a single recorder registers many co-occurring species at once. Yet automated recognition for insects lags far behind birds. The dominant deployed recogniser BirdNET (Kahl et al. 2021) classifies on the order of 35 insect species and the tropical-insect-PAM literature explicitly frames a scalable insect recogniser as the needed next step (Riede and Balakrishnan 2025).

The obstacle is recognition and specifically the data it requires. The standard route, training a bespoke classifier per project and fauna, depends on labelled reference audio, which does not exist for most insects. The great majority of insect diversity is undescribed or has never been recorded, so for exactly the taxa a survey most needs to detect there is often nothing to train on. Supervised recognition therefore presumes the knowledge that is missing, and the problem is sharpest in the species-rich tropical assemblages where monitoring would be most valuable (Riede and Balakrishnan 2025).

In the absence of per-species models, practitioners fall back on aggregate acoustic indices (such as the acoustic complexity index or the normalised difference soundscape index) that summarise soundscape energy as a proxy for biodiversity (Sueur et al. 2008; Pijanowski et al. 2011). These bypass the need for labels but are not species-resolved and are sensitive to recorder and processing choices. Unsupervised clustering can instead discover acoustic structure directly from recordings (Guerrero et al. 2023), but typically halts at unlabelled clusters, without connecting them to named taxa or a taxonomic frame.

Bioacoustic foundation models offer a third route. Models such as Perch (Merriënboer et al. 2025) and self-supervised encoders such as AVES (Hagiwara 2023) are trained on very large, weakly labelled corpora and yield transferable embeddings that generalise across taxa (Ghani et al. 2023; Stowell 2022). Audio-language variants extend to zero-shot, language-queried classification (Robinson et al. 2024, 2025). These are powerful but general-purpose systems. Yet none of these systems provide an insect-specific reference resource, open-set discovery of unnamed taxa, or a characterisation of where insect identification actually holds.

They also share a physical limit because they are pre-trained at low sample rates (Perch at 32 kHz) and hear nothing above roughly 16 kHz. This discards the ultrasonic band, where many katydids sing and where the entire echolocation repertoire of bats is produced.

EarthChirp is complementary and lightweight. It uses a frozen embedder with no training and targets insects specifically (as that is where the majority of open data is). It adds the open-set, noise-rejection and adaptation machinery a deployed monitoring tool needs. Here we ask what a frozen foundation model (FFM) can and cannot do for insect recognition, package the answer as a usable tool, and show that neither the approach nor its reach is confined to insects or to the audible band. Our contributions are four. (1) A 2,012-species insect acoustic reference library. (2) EarthChirp, a training-free recogniser that detects, names, discovers morphospecies, rejects anthropogenic noise, and adapts to local faunas. (3) A characterisation of where foundation-model insect identification is reliable and where it is not. (4) A demonstration that time-expansion lifts the frequency ceiling, carrying the pipeline across the ultrasonic spectrum to purely ultrasonic insects. We also test it on an unrelated ultrasonic taxon (bats), where it flags, clusters and coarsely places species the library has never seen. We do not claim to establish any new machine-learning methods or components (frozen-embedding transfer, prototype matching, few-shot adaptation) (Snell et al. 2017; Ghani et al. 2023). The contribution is the insect-specific resource, the integrated system, and the characterisation.

2. Materials and Methods

2.1 Embedding and the reference library

EarthChirp embeds 5-s audio windows with the frozen Perch 2.0 model (1536-d embeddings) and never retrains it. We use Perch 2.0 because it is among the strongest openly redistributable bioacoustic embedders on cross-taxon transfer benchmarks (Merriënboer et al. 2025; Ghani et al. 2023). On ECOSoundSet, an independent corpus that postdates both encoders, frozen Perch separates insects far better than the recent self-supervised AVES encoder (Hagiwara 2023) (Supplementary Figure S1). Because the pipeline is embedder-agnostic, a still-stronger frozen embedder would slot in unchanged. The reference library comprises species-, genus- and group-level prototypes, the L2-normalised mean embeddings of a taxon's reference clips. Reference recordings were harvested from the Global Biodiversity Information Facility (GBIF, Orthoptera + Cicadidae sound media) (GBIF.org 2026), Xeno-canto (the net-new species not already in GBIF) (Xeno-canto Foundation 2026), the Singing Insects of North America archive (SINA) (Walker, n.d.) and InsectSingers (Marshall and Cooley, n.d.), with per-species capping for class balance and species-name normalisation across sources.

2.2 The unified confidence gate

Our evaluations run on pre-segmented clips, with segmentation upstream of this recogniser, so by detection we mean the insect-versus-not discrimination step. What we characterise throughout is that recognition stage (detection, naming and morphospecies discovery). For each clip EarthChirp computes the cosine similarity to the nearest prototype. Above a confidence gate the clip is returned as a named species and below it the clip is routed to the morphospecies path, where unnamed clips are clustered (agglomerative, cosine). Each cluster is given a provisional

phylogenetic placement by its nearest genus/order prototype (the embedding space preserves coarse taxonomic structure). Named identification and morphospecies discovery are therefore two ends of one gate and not separate systems. For transparency, we report operating-point-dependent results as full sweeps rather than single tuned points. The morphospecies clustering quality is characterised across the whole agglomerative-cutoff range, and the confidence gate as a precision-coverage curve. For concreteness, we choose a fixed cosine-distance cutoff of 0.40 (cosine similarity 0.60), but the qualitative result holds across the sweep.

2.3 Non-insect outgroup and local adaptation

A non-insect outgroup (anthropogenic-noise and geophony prototypes) lets the recogniser reject non-insect audio by nearest-prototype rather than forcing every sound onto an insect. Local adaptation is training-free. A handful of labelled clips from the target soundscape rebuild local prototypes. We follow the prototypical few-shot paradigm (Snell et al. 2017) that underpins the DCASE few-shot bioacoustic benchmark (Nolasco et al. 2023). To recognise callers whose energy lies above the model's 16 kHz ceiling, we apply time-expansion (TE), a long-standing bat-detector technique (Jones et al. 2000). In TE, native-sample-rate audio is fed to the 32 kHz model as if recorded at 32 kHz, folding the full ultrasonic band into the model's input range (a translation of the spectrum, not retraining). The expansion factor is fixed by the recording's native sample rate (native rate divided by 32 kHz, so 3x for the 96 kHz katydids and 8x for the 256 kHz bats). This is not a tuned hyperparameter, and deployment uses the single fixed full-band factor. The dose-response across time-expansion factors in Section 3.6 varies it only as a diagnostic control, attributing the gain to genuinely supra-16 kHz content rather than to a windowing artefact. We use time-expansion as the simplest training-free option. A single time-expanded pass replaces the baseband pass (the audio resampled into the model's 16 kHz input band, discarding ultrasonic energy) and serves audible and ultrasonic callers alike (Section 3.6). Time-expansion also slows temporal modulation rates by the same factor, but the dose-response and unharmed-audible controls (Section 3.6) indicate the recognition gain is spectral rather than a modulation-rate artefact. Our European insect corpus contains only three purely ultrasonic species with sufficient material, too few to evaluate time-expansion, so Section 3.6 instead uses dedicated high-sample-rate ultrasonic-insect and bat corpora.

2.4 Evaluation data

We evaluate on independent corpora curated by other operators. ECOSoundSet (expert-annotated European Orthoptera + Cicadidae) (Funosas et al. 2026, 2025) is used for identification, morphospecies and adaptation. ESC-50 (Piczak 2015) is used for specificity against environmental and anthropogenic sound. BirdWeather is used for the deployment comparison against the incumbent. ECOSoundSet is independently curated and shares no recording identifier with our harvested library, so we cannot perform a programmatic recording-level de-duplication against the prototypes. Two facts bound any residual overlap. First, morphospecies discovery clusters ECOSoundSet embeddings among themselves and never consults the library, so prototype overlap cannot affect the clustering metrics. Second, cross-dataset naming against the library is low (≈26% open-set, 58% recording-disjoint few-shot), well below within-corpus accuracy and inconsistent with the inflation that gross prototype leakage would cause. A separate concern is whether the frozen Perch model saw these recordings during its own pretraining. ECOSoundSet combines

new fieldwork with recordings contributed by European specialists (Funosas et al. 2026), so incidental overlap with Perch's large training corpus cannot be fully excluded. The modest chance-corrected discovery scores (adjusted mutual information AMI 0.37, adjusted Rand index ARI 0.20), far from the near-perfect agreement that memorised embeddings would produce, argue against a material effect. For ultrasonic recognition (Section 3.6) we use dedicated high-sample-rate corpora. Barro Colorado Island katydids (96 kHz) are used for ultrasonic insects. To test cross-taxon generality, we use the NABat reference library of North American bats (256 kHz) (North American Bat Monitoring Program and U.S. Geological Survey 2022) as a known library with held-out novel species. We use MORCEGOTECA Neotropical bats (384 kHz) (Programa de Pesquisa em Biodiversidade and Instituto Nacional de Pesquisas da Amazônia 2026) as an out-of-distribution novelty probe. For an outgroup we use NABat's non-bat class. Open-set performance is summarised by reject-AUROC (area under the ROC curve, flagging species or sounds absent from the library), cluster NMI (normalised mutual information, recovering held-out species as morphospecies) and nearest-prototype taxonomic placement. Throughout the paper, evaluations are recording-disjoint wherever a corpus provides recording identifiers. The clips used to build or adapt a prototype and the clips used to test it come from different recordings, with one test clip per recording, so within-recording similarity cannot inflate a score.

3. Results

3.1 A global insect reference library

The library spans 2,012 species and 13,076 reference clips across Orthoptera and Cicadidae. EarthChirp embeds each clip with frozen Perch 2.0, and a single confidence gate returns a named species, an unnamed morphospecies, or a non-insect rejection (Figure 1a). Orthoptera dominate the library with 1,565 species against 447 cicadas, and Figure 1b shows the full taxonomic and source composition.

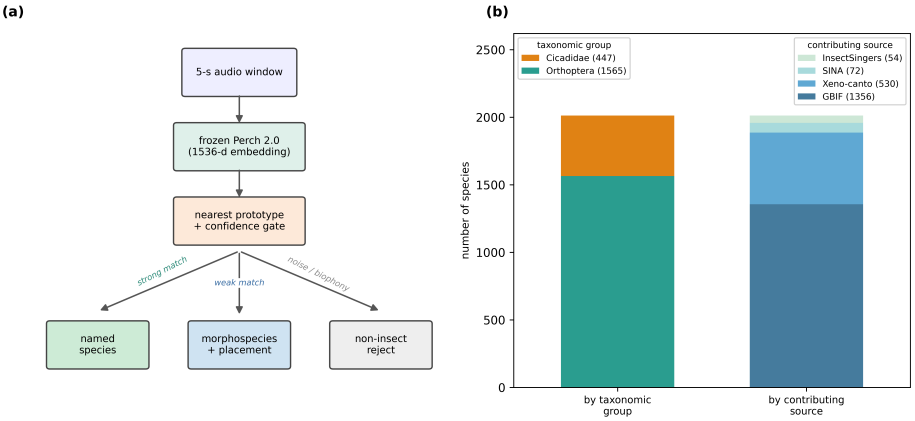

Figure 1. The EarthChirp recogniser (a) and reference-library composition (b). A 5-s window is embedded by frozen Perch 2.0; the nearest prototype and a confidence gate return a named species (strong match), a provisionally placed morphospecies (weak match), or a non-insect rejection. The library spans 2,012 species across Orthoptera and Cicadidae, assembled from GBIF, Xeno-canto, SINA and InsectSingers.

3.2 Detection, specificity, and the anthropogenic-noise failure mode

On ESC-50, the standard environmental-sound benchmark, EarthChirp separates insects from vertebrate biophony cleanly (AUROC 95% over 64 insect and 272 biophony source recordings). At one illustrative operating point on this curve, no bird, frog or mammal clip fires as an insect. ESC-50's insect positives are a narrow class, predominantly crickets, so we also characterise detection across the breadth of singing insects. On the 104-species ECOSoundSet, max-cosine to the insect prototypes separates insects from biophony (Supplementary Figure S2). Across the 103 species with at least three test clips, the median per-species AUROC is 99.9%, the lowest 87%, and pooled detection is 99% over 1,376 clips. The cricket-heavy ESC-50 set therefore understates the breadth rather than overstating it. The residual false-positive risk is anthropogenic mechanical noise (engines, machinery, AC hum), which is acoustically insect-like (continuous, tonal). Insect-versus-mechanical AUROC is 74% (64 insect and 569 mechanical source recordings). Anthropogenic sound is a recognised confound in insect bioacoustics (buzzdetect, for example, treats aircraft, sirens and ambient noise as negative classes for pollinator-flight detection (Hearon et al. 2025)). A training-free anthropogenic-noise + geophony outgroup prototype raises discrimination to 93%. It is built from a held-out half of the noise and geophony recordings and scored on the disjoint remaining half. At a retention-matched operating point it keeps 85% of true insect detections (Figure 2). These AUROCs are threshold-free, summarising performance across all operating points. The illustrative figures we quote, the no-false-fire example and the 85% retention, are single points on those curves, not an operating point tuned on the evaluation data.

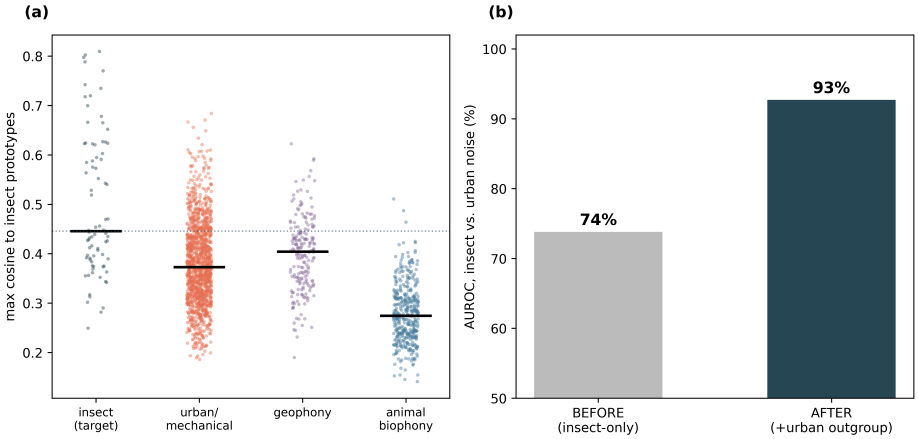

Figure 2. The residual false-positive risk is anthropogenic mechanical noise rather than animal biophony, and a training-free outgroup prototype lifts insect-vs-noise discrimination. (a) On ESC-50, insect-vs-biophony separates cleanly while urban/mechanical noise overlaps insects. (b) Insect-vs-urban discrimination rises from AUROC 74% (insect-only) to 93% with a non-insect outgroup prototype, retaining 85% of true insect detections.

3.3 Zero-shot morphospecies discovery

On the full independent ECOSoundSet (104 European Orthoptera and Cicadidae species recorded by other operators), EarthChirp's agglomerative clustering of insect embeddings groups clips into species-pure morphospecies without labels. Cluster purity, the fraction of each cluster belonging to its dominant species, is 0.79. A morphospecies is an unnamed cluster that resolves clips at roughly the species grain,

not the coarser genus. Giving it a correct taxonomic name is a separate and harder problem, and that name is trustworthy only at the broad group rank (genus is rough, and species-level naming is deferred to the few-shot adaptation of Section 3.4). The clustering over-segments about six-fold (636 clusters versus 104 species at the 0.40 cutoff), so the chance-corrected agreement is the honest measure of recovered structure (AMI 0.37 and ARI 0.20). The high NMI (0.77) is partly inflated by that same over-segmentation. This still confirms that EarthChirp's detection-and-discovery generalises off the model's training distribution, if more modestly than NMI alone would suggest. Species-pure over-segmentation is the acceptable behaviour for morphospecies discovery, whose goal is coherent unnamed acoustic units rather than an exact species count. To avoid pseudoreplication these metrics use one clip per recording-species, so near-duplicate clips from the same recording cannot inflate the agreement. They are reported at a single fixed cosine-distance cutoff of 0.40 rather than each at its own best threshold, with 95% intervals from a leave-one-species-out jackknife, and Figure 3b shows the full sweep. Because over-segmentation biases the cluster count upward, the number of morphospecies should be read as an upper bound on richness rather than a calibrated diversity figure. We also reproduced the prevailing feature-based representation (Guerrero et al. 2023) (frequency descriptors plus linear-scale cepstral coefficients) on the same clips. The frozen embedding clusters insects at least as well as these hand-crafted features (NMI 0.78 versus 0.76, each at its own best threshold), and fusing the two adds nothing.

Placing each discovered cluster on the taxonomy by its nearest reference prototype is the weakest and most uneven step, and we report it carefully because it supports our claim that only coarse taxonomic placement survives the domain shift. At the coarse group level (Orthoptera, an order, versus the cicadas, family Cicadidae, so the two groups sit at different taxonomic ranks) raw accuracy is 89%, but the test set is 80% Orthoptera (a majority-class guesser would score ~80%), so the honest figure is a balanced accuracy of 74% (Cohen's κ 0.58). Orthoptera recall is near-perfect (99%) while Cicadidae recall is only near chance (48%). Genus placement, against the full 619-genus candidate library, is rougher still (30%, Cohen's κ 0.29). The bat genus task in Section 3.6 reaches κ 0.50, but the two are not directly comparable. Cohen's κ is not invariant to candidate-set size, and the two candidate sets differ roughly 80-fold (619 genera versus 8). Restricting the insect candidates to the 56 genera actually present in the corpus, a set closer in size to the bat task, lifts genus placement to 43% (κ 0.41), and the correct genus lies within the top five of the full 619 for 55% of clips. The gap to the bat number is therefore candidate-set size together with cross-dataset domain shift (European clips matched against prototypes built mostly from non-European species in the same genus) rather than an insect-specific failure. Taxonomic placement of discovered clusters is therefore indicative rather than definitive, and reliable only for the dominant group (Orthoptera). The morphotype itself survives the domain shift, while its taxonomic label does not.

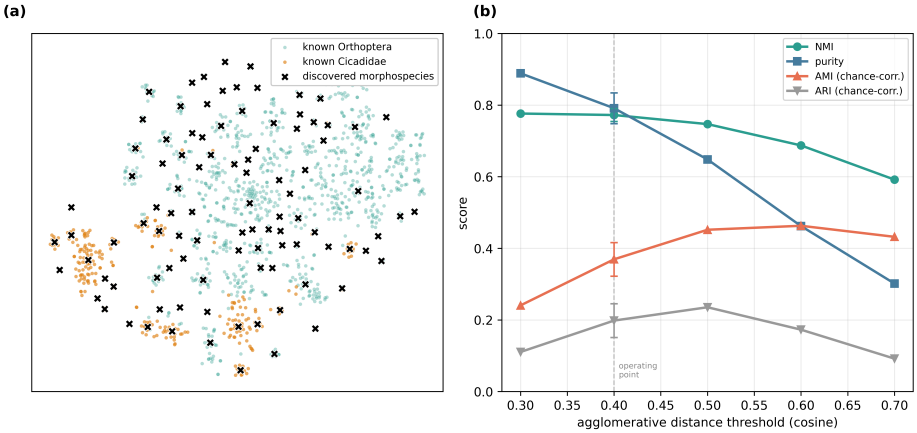

Figure 3. Zero-shot morphospecies discovery on independent, expert-labelled ECOSoundSet (104 species, one clip per recording-species to avoid pseudoreplication). (a) t-distributed stochastic neighbour embedding (t-SNE) of known-species means (coloured by group, Orthoptera versus the cicadas) with the discovered morphospecies centroids (black crosses), the discovered clusters fall among the known species more cleanly for Orthoptera than for the minority cicadas (Cicadidae). (b) Cluster quality across agglomerative thresholds, with 95% jackknife CIs at the 0.40 operating point: the chance-corrected AMI and ARI are the honest measure (purity stays high under species-pure over-segmentation, which also inflates NMI). The species-coloured clip t-SNE, a given for any capable foundation model, is shown in Supplementary Figure S3.

3.4 Cross-dataset identification and few-shot adaptation

Fine species naming does not transfer zero-shot to independent data. On the expert-labelled ECOSoundSet corpus (104 European Orthoptera and Cicadidae species recorded by other operators), cross-dataset open-set accuracy (naming against the full 2,012-species library) is ≈26%, and a learned classifier head does no better (≈25%). This confirms the limit is the embedding's domain shift rather than the matching rule. In realistic closed-set deployment over this local fauna, library prototypes already reach 44%. Few-shot local adaptation recovers to 58% species / 64% genus with ten reference clips per species (mean ± standard deviation over 5 seeds, Figure 4a). The split is fully recording-disjoint, with the adaptation and test clips drawn from different recordings of each species, so the recovery is not inflated by within-recording similarity. Tested out to forty reference clips, naming plateaus by about ten (Figure 4a), so the ceiling is the frozen embedding rather than reference-data quantity. A single clip is noisier than the many-clip library prototype and slightly hurts. What governs accuracy is recording-condition match. Naming reaches ~65% when a reference clip comes from the test recording's own site and falls to ~39% for a novel site, with ~58% pooled over all sites. The novel-site loss is roughly half recording equipment (matched recorders recover 44% versus 27%) and half genuine between-site variation, so on-site reference recordings are the practical lever. A logistic-regression classifier trained on the same K reference clips per species does no better than the training-free prototype (56.6% versus 58.0% species top-1 at ten clips, and on par at every K). The no-training recipe therefore sacrifices no accuracy to a small supervised model fitted on the same local data. Under the unified gate, naming precision on the confident subset rises with the confidence threshold. With ten local clips it reaches about 71% over the most confident fifth of clips, and about 69% over the most confident half, while the low-confidence

remainder is correctly routed to morphospecies (Figure 4b). Stated plainly, the usable naming operating point names only the most confident ~20% of clips, of which about seven in ten are correct. The rest are better surfaced as morphospecies than named. EarthChirp is therefore a coarse classifier and discovery tool whose naming sharpens with on-site reference clips but saturates short of fine identification, not a precise zero-shot namer. Augmentation-averaging the embeddings did not improve the zero-shot floor.

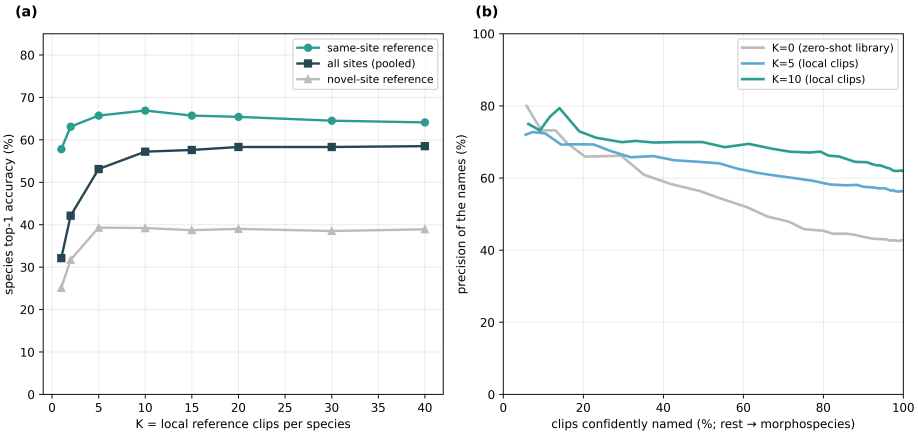

Figure 4. Cross-dataset identification and few-shot adaptation. (a) Few-shot species naming saturates by about ten reference clips and is governed by recording-condition match rather than clip count. It reaches ~65% when a reference clip shares the test recording's site and ~39% for a novel site, ~58% pooled over all sites (104 ECOSoundSet species, closed-set, recording-disjoint, mean over 5 seeds). The plateau places the limit on the frozen embedding, not local-data quantity. (b) The unified confidence gate, recording-disjoint naming precision versus coverage. Zero-shot the gate is weak and most clips route to morphospecies, but a few local clips sharpen it to about 71% precision over the most confident fifth of clips (69% over the most confident half), while clips below the gate become morphospecies.

3.5 Deployment comparison against the incumbent

On BirdWeather, EarthChirp recovers insect signal the deployed incumbent discards. It recalls 100% of BirdNET's own insect detections and flags insects in 63% of clips BirdNET could only label "bird", at a 5% false-positive rate on off-season controls (Figure 5). These percentages are an indicative deployment signal, not a measured identification rate. The 63% and 5% are uncalibrated because BirdNET's insect vocabulary is too small to serve as ground truth, so we present the comparison as deployment motivation rather than a benchmark.

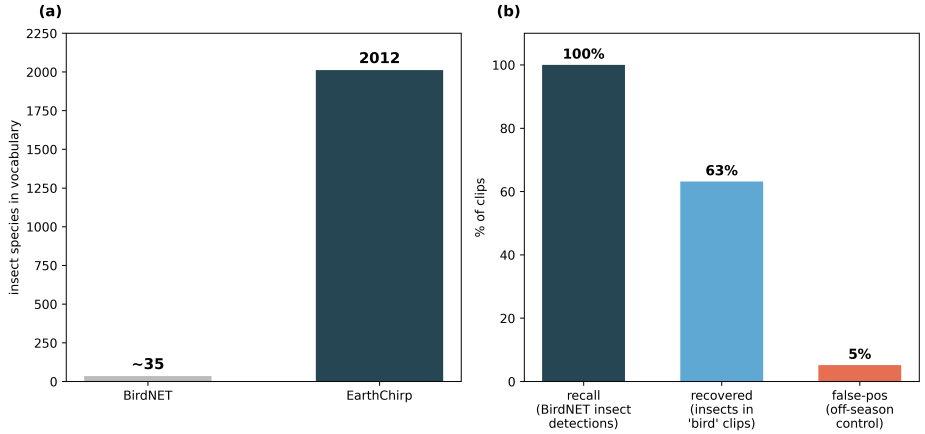

Figure 5. Deployment comparison against the incumbent on BirdWeather. (a) Insect-species coverage: BirdNET classifies ~35 insect species, EarthChirp 2,012. (b) EarthChirp recalls 100% of BirdNET's own insect detections and flags insects in 63% of clips BirdNET binned as "bird", at a 5% false-positive rate on off-season controls.

3.6 Time-expansion unlocks ultrasonic recognition and generalises to bats

The 16 kHz ceiling is a fixed limit on what the frozen model can hear, but not on what the pipeline can recognise. Time-expansion folds ultrasonic energy down into the model's band, letting the same frozen embedder read it. A recent benchmark independently finds time-expansion the strongest method for the purely ultrasonic case (Sarkar et al. 2026). That study is supervised and closed-set, and its own proposed method is a learned multi-band encoding rather than time-expansion. We use it for a different reason. Because time-expansion alters only the input, EarthChirp's frozen, training-free pipeline carries over unchanged into open-set discovery and onto a new taxon. A single time-expanded pass improves recording-disjoint species discrimination on both a focal and a soundscape recording set of 96 kHz Barro Colorado Island katydids. On the focal set, the ultrasonic-call subset rises from 84% to 98% (nearest-centroid accuracy over 22 species and 832 recordings, chance ~5%). On the harder, multi-species soundscape set, the same subset shows a dose-response as time-expansion folds progressively more of the band, climbing from 21% at baseband through 30% at x2 (0 to 32 kHz) to 44% at x3 (0 to 48 kHz) (9 morphotypes, 25 recordings, chance ~11%). The soundscape endpoints are underpowered. Taken alone, the baseband-to-x3 jump on the soundscape set is not statistically significant, since the small-sample intervals overlap. The case rests instead on the steady monotonic rise across the three factors, together with the focal-set gain, which is significant and disappears once the ultrasonic band is removed. Together, these localise the improvement to genuinely supra-16 kHz content rather than a windowing artefact. Time-expansion also does not harm audible callers (which pass through the same window unchanged), so one pass serves both (Figure 6a). This delivers the time-expansion evaluation that our European insect corpus, with only three purely ultrasonic species, was too small to support (Section 2.3).

The same frozen pipeline then carries over unchanged to a taxon the embedder never trained on. Using the voucher-verified NABat reference library of North American bats (256 kHz) as a known library with held-out novel species, frozen-Perch-on-time-expanded audio performs open-set discovery (Figure 6b, Supplementary Table S1). Against 355 held-out known-species clips, each from a distinct recording, it reliably rejects non-bat noise (reject-AUROC 99%, 45 recordings) and out-of-region

Neotropical bats (≈100%, 17 recordings), and rejects held-out North American species more moderately (75%, 445 recordings). Within a well-sampled fauna a held-out species resembles its known congeners. It then recovers those held-out species as coherent morphospecies (chance-corrected cluster ARI 0.53 to 0.58, NMI 0.71 to 0.88) and places them taxonomically (genus 68%, Cohen's κ 0.50 over an 8-genus candidate set, for novel clips against a 77% known-species control). A clip can only be placed into a genus the library has already seen, so this genus score is computed over held-out species whose genus is represented in the known set. Species from entirely novel genera cannot be placed and are excluded. Bat calls lie almost entirely above 16 kHz, so the baseband model is near-blind, with noise rejection at 58%, genus placement at 26%, and chance-corrected cluster ARI near zero despite a superficially non-zero NMI (Supplementary Table S1). On the limited bat audio currently available, the result indicates that EarthChirp's pipeline can carry across the ultrasonic spectrum.

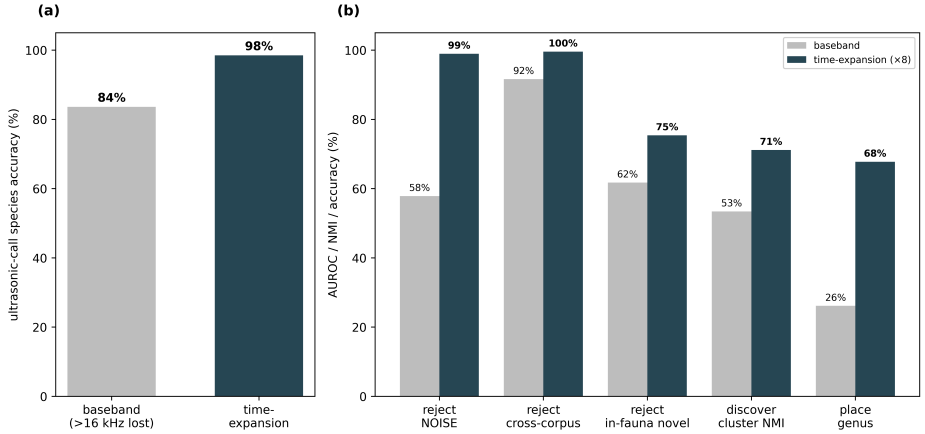

Figure 6. Time-expansion extends EarthChirp across the ultrasonic spectrum. (a) On 96 kHz Barro Colorado Island katydids, a single time-expanded pass lifts recording-disjoint species discrimination of ultrasonic callers from 84% (baseband, blind above 16 kHz) to 98%, without harming audible callers. (b) The same frozen pipeline generalises to bats (NABat, 256 kHz). With time-expansion it rejects non-bat noise (AUROC 99%) and out-of-region Neotropical species (≈100%), and rejects a held-out species within the same fauna more moderately (75%, the ecologically common case), discovers held-out species as morphospecies (chance-corrected ARI 0.53 to 0.58, NMI 0.71 to 0.88) and places them by genus (68% for novel clips, Cohen's κ 0.50 over 8 candidate genera, against a 77% known-species control), while the baseband model is near-blind.

4. Discussion

EarthChirp is a training-free insect detector, coarse classifier and morphospecies-discovery tool that is robust to anthropogenic noise and adaptable to local faunas, rather than a precise zero-shot cross-region species namer. What survives the embedding's domain shift splits cleanly. Coarse acoustic structure (insect versus not, group, and morphotype) transfers to independent recordings, while fine between-species structure does not. A frozen model trained on continental-scale bird and general bioacoustic audio (Merriënboer et al. 2025; Ghani et al. 2023) therefore

already encodes enough acoustic regularity to register that a singing insect is present and to recover which discovered clips belong together, but not to say reliably which named species produced a call recorded under conditions it has never seen.

The practical consequence is a deployment recipe in which no model is ever trained. Detection and discovery run anywhere with no local data, so a project can survey a new soundscape, separate insect sound from everything else, and surface coherent morphospecies immediately. Species naming is an optional upgrade that asks only for a handful of on-site reference clips, which are averaged into prototypes (Snell et al. 2017) and added to the library in seconds rather than learned through gradient descent. In practice, with ten local reference clips per species, the recogniser returns a species name for the most confident ~20% of clips, correct about 71% of the time, and surfaces the rest as morphospecies. The distinction matters operationally. The per-site cost of a bespoke classifier is data collection, labelling and training, whereas EarthChirp's per-site cost is at most a few exemplar recordings, and is zero wherever detection and discovery alone suffice. The few-shot step is the standard prototypical paradigm behind the DCASE bioacoustic benchmark (Nolasco et al. 2023), applied here without retraining the embedder.

EarthChirp's failure mode is more specific than false positives in general. Other animals are not the problem, since, on independent environmental audio, it cleanly separates insects from vertebrate biophony. The binding constraint is anthropogenic mechanical noise, which is continuous and tonal in the same way insect song is and overlaps insect embeddings far more than birdsong or frog chorus. This is consequential because real monitoring sites are rarely free of engines, machinery and electrical hum, and an unguarded detector will convert that hum into spurious insect activity. To mitigate this, EarthChirp uses a training-free non-insect outgroup prototype rather than forcing every sound onto an insect, which sharply improves insect-versus-noise discrimination while keeping most true detections. Anthropogenic sound is a recognised confound in insect bioacoustics, handled in related single-purpose settings such as separating insect chorus from environmental noise (Brown et al. 2019) and detecting pollinator flight against background (Hearon et al. 2025). But its role as the binding false-positive constraint for foundation-model insect recognition, and a training-free remedy for it, have not previously been characterised.

For discovery, the clustering algorithm is unchanged from prior work. What matters is where the clusters live. Unsupervised acoustic clustering for species discovery is established, most directly in the feature-based approach that segments calls and clusters hand-crafted frequency and cepstral descriptors (Guerrero et al. 2023, 2025). On the same recordings, FFM embeddings cluster insects no better than those descriptors, and fusing the two adds nothing. The embedding instead lives in a space shared with the reference library, so each discovered cluster can be placed against named taxa automatically rather than left as an unlabelled sonotype for an expert to associate with a species at each site, the step that this prior work itself identifies as open future work (Guerrero et al. 2025). Placement is the weaker, more uneven step. For insects it works for the dominant group but is asymmetric (balanced accuracy 74% at the coarse group level, near-chance recall for the minority family, Cicadidae) and rough at genus, so the discovered morphotype (the acoustic cluster itself) is the durable unit, while the taxonomic name attached to it remains only indicative.

The model's 16 kHz ceiling, which a practitioner might take as a hard limit on what a 32 kHz model can hear, is instead a property of the raw input that time-expansion removes (Jones et al. 2000; Sarkar et al. 2026). Folding the ultrasonic band into the model's range recovers purely ultrasonic insects and carries the entire pipeline

unchanged to bats, a taxon the embedder was never trained on. With the limited open bat audio currently available, it rejects non-bat noise and out-of-region species, clusters held-out species into coherent morphospecies, and provisionally places them by genus above chance. The broader implication is that none of EarthChirp's machinery is specific to insects or to the audible band. The reference library is insect-specific, but the recipe (a frozen embedder, a tiered prototype library, and an open-set gate) carries across the ultrasonic spectrum. We present the bat result as a generality demonstration rather than a bat-recognition benchmark.

Several limitations bound these claims. Every quantitative naming and discovery result here is measured on the 104-species European ECOSoundSet. The 2,012-species library appears only as the open-set candidate pool, so its size reflects the recognition breadth the method can address, not a validated identification breadth. The zero-shot naming floor is set by the data available to build prototypes, and it is lowest exactly where it would be most useful, in the species-rich, under-recorded tropical assemblages that motivate the work (Riede and Balakrishnan 2025). The same scarcity skews the geographic coverage of the reference library. Its sources are temperate-heavy (SINA and InsectSingers are North American, Xeno-canto and GBIF are global but temperate-skewed), so coverage is densest in the Holarctic and sparse in exactly the tropical regions the work targets. For ultrasonic taxa the data problem is sharper. Openly licensed, full-spectrum, verified bat audio is genuinely scarce (Supplementary Appendix S1), so the cross-taxon probe rests on a small Neotropical sample and a single regional reference library, and the bat results should be read accordingly. Open-set rejection is also not uniform. It is strong for genuinely out-of-distribution sound such as noise and out-of-region taxa, but only moderate for a held-out species within a well-sampled fauna, because such a species closely resembles its known congeners. Finally, fine identification still depends on per-fauna reference clips, and the automatic placement of discovered insect clusters is reliable only at coarse taxonomic ranks.

These limitations share a direction. EarthChirp is a reference-library method whose detection, discovery, naming and rejection all improve directly with the breadth of the library when users add their own references. A practitioner extends the tool to a new fauna, region or taxon by embedding reference recordings they are licensed to use, including restricted-licence libraries that cannot be redistributed but can be embedded locally. The clearest paths forward are therefore data rather than architecture, namely wider and more geographically even insect coverage, and openly licensed ultrasonic reference sets for the taxa that currently lack them. Additionally useful would be addressing the 16 kHz ceiling, which limits resolution, with embedders trained natively at higher sample rates (Sarkar et al. 2026). As these improvements accumulate, the same frozen, training-free recipe should reach more of the acoustic world.

Data and code availability

The EarthChirp reference library (prototypes) and classifier head (~29 MB) are archived on Zenodo (DOI 10.5281/zenodo.20696319), and all analysis code is at https://github.com/foundfold/earthchirp. Reference recordings derive from GBIF (GBIF.org 2026), Xeno-canto (Xeno-canto Foundation 2026), SINA (Walker, n.d.) and InsectSingers (Marshall and Cooley, n.d.). Evaluation uses ECOSoundSet (Funosas et al. 2025), ESC-50 (Piczak 2015) and BirdWeather. No reference audio is redistributed. Only derived embeddings and source attributions are released.

Author contributions

M.M. conceived the study, developed EarthChirp, performed all analyses, and wrote the manuscript.

Conflict of interest

The author declares no competing interests.

Acknowledgements

This work uses open data and tools from GBIF, Xeno-canto, the Singing Insects of North America archive, InsectSingers, ECOSoundSet, ESC-50, BirdWeather and the open Perch 2.0 model. We thank their contributors and maintainers.

References

Brown, Alexander, Saurabh Garg, and James Montgomery. 2019. “Automatic Rain and Cicada Chorus Filtering of Bird Acoustic Data.” *Applied Soft Computing* 81: 105501. https://doi.org/10.1016/j.asoc.2019.105501.

Funosas, David, Elodie Massol, Yves Bas, et al. 2025. *ECOSoundSet: Acoustic Database and Annotations*. Zenodo. https://doi.org/10.5281/zenodo.15043892.

Funosas, David, Elodie Massol, Yves Bas, et al. 2026. “A Finely Annotated Dataset for the Automated Acoustic Identification of European Orthoptera and Cicadidae.” *Scientific Data* 13: 830. https://doi.org/10.1038/s41597-026-07150-1.

GBIF.org. 2026. *GBIF Occurrence Download (Orthoptera and Cicadidae Sound Media)*. GBIF Secretariat.

Ghani, Burooj, Tom Denton, Stefan Kahl, and Holger Klinck. 2023. “Global Birdsong Embeddings Enable Superior Transfer Learning for Bioacoustic Classification.” *Scientific Reports* 13: 22876. https://doi.org/10.1038/s41598-023-49989-z.

Gibb, Rory, Ella Browning, Paul Glover-Kapfer, and Kate E. Jones. 2019. “Emerging Opportunities and Challenges for Passive Acoustics in Ecological Assessment and Monitoring.” *Methods in Ecology and Evolution* 10 (2): 169–85. https://doi.org/10.1111/2041-210X.13101.

Guerrero, Maria J., Carol L. Bedoya, José D. López, Juan M. Daza, and Claudia Isaza. 2023. “Acoustic Animal Identification Using Unsupervised Learning.” *Methods in Ecology and Evolution* 14 (6): 1500–1514. https://doi.org/10.1111/2041-210X.14103.

Guerrero, Maria J., Camilo Sánchez-Giraldo, César A. Uribe, Víctor M. Martínez-Arias, and Claudia Isaza. 2025. “Graphical Representation of Landscape Heterogeneity Identification Through Unsupervised Acoustic Analysis.” *Methods in Ecology and Evolution* 16 (7): 1255–72. https://doi.org/10.1111/2041-210X.70041.

Hagiwara, Masato. 2023. “AVES: Animal Vocalization Encoder Based on Self-Supervision.” *ICASSP 2023 – IEEE International Conference on Acoustics,*

Speech and Signal Processing, 1–5. https://doi.org/10.1109/ICASSP49357.2023.10095642.

Hearon, Luke E., Lillian H. P. Johnson, James Underwood, Chia-Hua Lin, and Reed M. Johnson. 2025. "Buzzdetect: An Open-Source Deep Learning Tool for Automated Bioacoustic Pollinator Monitoring." *Journal of Insect Science* 25 (6): ieaf104. https://doi.org/10.1093/jisesa/ieaf104.

Jones, Gareth, Nancy Vaughan, and Stuart Parsons. 2000. "Acoustic Identification of Bats from Directly Sampled and Time Expanded Recordings of Vocalizations." *Acta Chiropterologica* 2 (2): 155–70.

Kahl, Stefan, Connor M. Wood, Maximilian Eibl, and Holger Klinck. 2021. "BirdNET: A Deep Learning Solution for Avian Diversity Monitoring." *Ecological Informatics* 61: 101236. https://doi.org/10.1016/j.ecoinf.2021.101236.

Marshall, David C., and John R. Cooley. n.d. *InsectSingers: Cicada Song Recordings*. Https://www.insectsingers.com.

Merriënboer, Bart van, Vincent Dumoulin, Jenny Hamer, Lauren Harrell, Andrea Burns, and Tom Denton. 2025. *Perch 2.0: The Bittern Lesson for Bioacoustics*. https://doi.org/10.48550/arXiv.2508.04665.

Nolasco, Inês, Shubhr Singh, Veronica Morfi, et al. 2023. "Learning to Detect an Animal Sound from Five Examples." *Ecological Informatics* 77: 102258. https://doi.org/10.1016/j.ecoinf.2023.102258.

North American Bat Monitoring Program, and U.S. Geological Survey. 2022. *Training Dataset for NABat Machine Learning V1.0*. U.S. Geological Survey. https://doi.org/10.5066/P969TX8F.

Piczak, Karol J. 2015. "ESC: Dataset for Environmental Sound Classification." *Proceedings of the 23rd ACM International Conference on Multimedia*, 1015–18. https://doi.org/10.1145/2733373.2806390.

Pijanowski, Bryan C., Luis J. Villanueva-Rivera, Sarah L. Dumyahn, et al. 2011. "Soundscape Ecology: The Science of Sound in the Landscape." *BioScience* 61 (3): 203–16. https://doi.org/10.1525/bio.2011.61.3.6.

Programa de Pesquisa em Biodiversidade, and Instituto Nacional de Pesquisas da Amazônia. 2026. *Morcegoteca: A Virtual Library of Bats' Ultrasounds*. Https://ppbio.inpa.gov.br/en/Bat_Library.

Riede, Klaus, and Rohini Balakrishnan. 2025. "Acoustic Monitoring for Tropical Insect Conservation." *Philosophical Transactions of the Royal Society B* 380 (1928): 20240046. https://doi.org/10.1098/rstb.2024.0046.

Robinson, David, Marius Miron, Masato Hagiwara, et al. 2025. "NatureLM-audio: An Audio-Language Foundation Model for Bioacoustics." *International Conference on Learning Representations (ICLR)*.

Robinson, David, Adelaide Robinson, and Lily Akrapongpisak. 2024. "Transferable Models for Bioacoustics with Human Language Supervision." *ICASSP 2024 – IEEE International Conference on Acoustics, Speech and Signal Processing*, 1–5. https://doi.org/10.1109/ICASSP48485.2024.10447250.

Sarkar, Eklavya, Marius Miron, David Robinson, et al. 2026. *Beyond the Baseband: Adaptive Multi-Band Encoding for Full-Spectrum Bioacoustics Classification*. https://doi.org/10.48550/arXiv.2604.27936.

Snell, Jake, Kevin Swersky, and Richard S. Zemel. 2017. "Prototypical Networks for Few-Shot Learning." *Advances in Neural Information Processing Systems (NeurIPS)*.

Stowell, Dan. 2022. "Computational Bioacoustics with Deep Learning: A Review and Roadmap." *PeerJ* 10: e13152. https://doi.org/10.7717/peerj.13152.

Sueur, Jérôme, Sandrine Pavoine, Olivier Hamerlynck, and Stéphanie Duvail. 2008. “Rapid Acoustic Survey for Biodiversity Appraisal.” *PLOS ONE* 3 (12): e4065. https://doi.org/10.1371/journal.pone.0004065.

Sugai, Larissa S. M., Thiago S. F. Silva, José W. Ribeiro Jr., and Diego Llusia. 2019. “Terrestrial Passive Acoustic Monitoring: Review and Perspectives.” *BioScience* 69 (1): 15–25. https://doi.org/10.1093/biosci/biy147.

Walker, Thomas J. n.d. *Singing Insects of North America (SINA)*. Zenodo. https://doi.org/10.5281/zenodo.13312646.

Xeno-canto Foundation. 2026. *Xeno-Canto: Sharing Animal Sounds from Around the World*. Https://xeno-canto.org.