Lyrebird: a physically interpretable syllable inventory for 2 650 wild bird species

Mag. art. Shahab Nedaei
variable.gallery
ned.tabulov@gmail.com

4 August 2026

Abstract

Lyrebird measures syllables in wild bird recordings from xeno-canto and iNaturalist. It fits the Laje–Mindlin normal-form syrinx oscillator to each measured syllable type by coordinate descent. It learns a per-species type inventory and a first-order grammar over that inventory. It ships the result as a browser application, where live or synthesised sound lights the acoustically nearest regions of a three-dimensional map. The corpus holds 378 224 measured syllables from 11 687 recordings. The fit covers 2 650 species and 16 283 syllable types. The pipeline decodes each recording, describes it as numbers, and deletes it. No audio is redistributed and no model weights reach the browser. Every claim the interface makes carries one of five evidence labels. Four things in this corpus are measured and hold up. First, the learned tokens behave as units: a first-order transition model buys 0.98 bits per token on held-out bouts and *costs* 0.75 bits on a within-species shuffle of the same tokens. An unbiased estimator that pays for parameters it cannot use is the clean form of that claim. Second, the 32-dimensional embedding carries real per-species geometry: over 12 736 held-out trials, a species' own types find the held-out type at top-1 0.540 against a chance rate of 0.00007. A shuffled control sits at chance. Third, repeat classes take the corpus from 48 to 98 species with a measured song, and the permutation control stays at 0.170 % of candidate bouts. Fourth, the physical model reproduces the median measured type at a 28-dimensional descriptor cosine of 0.699. Four results run against the design, and we report them at the same weight. The scalar the pipeline optimises is a cosine in a 28-dimensional descriptor space, and over 1 200 fitted types it is *positively* correlated with audible spectral distortion — Spearman +0.200 against mel-cepstral distortion at $p = 3\times10^{-12}$, and +0.186 against multi-resolution spectral error, two metrics that agree with each other at 0.925. Higher cosine goes with a worse sounding fit. Second, that is not hypothetical, and it is not uniform either: of the two cosine-gated adoptions that can be re-judged without refitting, the respiratory block gained +0.034 of median cosine and *cost* 4.18 dB of median distortion, improving only 33 % of 175 types, while the two-voice block gained +0.071 and *won* 1.13 dB on 55.6 % of 1 737 types. A proxy anti-correlated with fidelity across types removes the guarantee that a gate built on it is safe; it does not make every such gate harmful. Third, the learned embedding organises by recordist before it organises by bird, at a nearest-neighbour lift of +0.488 against +0.196, and four independent attempts to remove that confound failed — two of them using archive-scale external models, which disagree with each other by more than either differs from ours. That confound has a threshold: the cost of blocking an evaluation split by recordist falls from 30.8 points of top-1 at two recordists per species to 3.1 at six or more, and the *same individual* recordists are the identifiable ones in Perch and BirdNET alike ($r = +0.85$ partial). Fourth, with every type granted its own cluster, 0.602 of syllables have a nearest centroid belonging to their own type. That number is the ceiling on "sing and the right region lights". Appendix A lists all 58 pre-registered gates and their outcomes. 20 passed, 37 ended in a written refusal, and 1 is unrun. The refusals are as much of the contribution as the system is.

Keywords: bioacoustics; birdsong; syrinx model; gesture inversion; contrastive embeddings; sequence statistics; negative results; pre-registration.

1 Premise

Lyrebird is an instrument in three parts.

It *sings* from physics. Every note is integrated at audio rate from the low-dimensional normal-form oscillator of the Laje–Mindlin lineage [24, 40, 42, 46], driven by the two scalar gestures a real bird controls: air-sac pressure and syringeal tension. Nothing is sampled, no waveform is replayed, and no neural generator is involved. The parameters that produce a given note are the parameters a physiologist would name.

It *remembers* without keeping audio. The harvester fetches a recording, decodes it, segments it into syllables, describes each syllable as a 28-dimensional acoustic descriptor and a 32×23 log-mel patch, and deletes the file. The production store holds 378 224 such measurements over 11 687 recordings and 3 812 species, of which 2 650 species carry a fitted inventory of 16 283 syllable types. The measurements are not invertible to recognisable audio, so the store is a description of an archive, not a copy of one.

It *shows* the result as a map. One point is one measured syllable. Points are placed by projecting the space in which syllable identity is actually decided. A learned 32-dimensional embedding. Down to three dimensions,

and clusters in that space are labelled with the syllable type, the documented behavioural function where one exists, and the citation for it. When a microphone hears a sound, or when the synthesiser sings one, the regions of the map nearest to it in the acoustic space light up.

What the system claims about meaning is narrow and is printed on its face. It maps a call type to a documented behavioural function, and it attaches to every phrase it speaks one of five labels: *experimental* (function established by playback experiment), *documented* (function described in the literature), *inferred* (measured in this corpus, function not established), *invented* (a designed mapping, named as designed), and *unlabelled* (real acoustics, no documented function). Playback is the only proof of function the field accepts [65]; Lyrebird has run none of its own, so everything past the literature is labelled as inference or as invention.

This paper reports the instrument and, at greater length, the measurements that constrain it. The project's working discipline is that a change to fitting, to an objective or to a space is judged by a metric written down before the run, on a fixed corpus, and that a result which contradicts the plan is recorded in place, not deleted. Fifty-eight such gates have been run. Thirty-six of them failed. Sections 8 to 6 are organised around what survived and what the failures rule out.

2 Motivation

The gap the project was built into

The immediate provocation was a piece of viral content: an unindexed reel captioned *AI is decoding what birds have been saying all along*, composited from three real things. The media cycle around machine learning for animal communication, a photograph of a nightingale, and a Sainburg-style two-dimensional embedding of a vocal repertoire [66]. The composite implies a translator. None of its three ingredients is one.

Every ingredient, however, is buildable, and the honest version of the promise is worth building. Repertoire characterisation works. Functional decoding. Linking a signal to a context or a behaviour. Works for the handful of systems where playback experiments exist. Recovering propositional content does not work and is not close [65]. There is also a structural asymmetry that any such system runs into: birdsong carries rich syntax and weak or absent reference, while calls carry reference and little syntax [54]. A system that conflates the two produces the reel.

The interesting question is therefore not "can a model translate birdsong" but *what can a large archive of wild recordings actually license*, and what does it cost to say so precisely. That is the question this work answers, mostly in the negative, and the negative answers are specific enough to be useful.

Why a physical model and not a generative one

Bird-vocalisation synthesis has a validated mechanistic option, which is unusual for a biological signal. The normal-form oscillator reproduces canary song from two scalar gesture curves [24]; the bird's own premotor neurons encode those same gestures [3]; the model runs as a real-time neural prosthesis [4]; and, in the strongest available test, 256 playback trials on 26 wild rufous-collared sparrows at 15 sites found no significant difference between responses to real song and to song synthesised from the model [7]. Heterospecific song evoked no response at all in the same design.

A neural audio generator would have none of that. It also could not carry the evidence labels, because a label has to be grounded in a mechanism a reader can inspect, and a generator's mechanism is a weight matrix. The decision to keep the synthesiser physical is therefore not a tooling preference; it is what makes the honesty rule enforceable.

3 Related work

3.1 Physical models of the avian syrinx

The lineage this work builds on treats the syrinx as a low-dimensional nonlinear oscillator, not as a signal generator to be reverse-engineered. Measurements on the excised syrinx show period doubling, mode locking and chaos, so the source is not a whistle [20]. The minimal dynamical system reproducing that behaviour sits at the Takens–Bogdanov point of the labial model, where a Hopf line is tangent to a saddle-node curve [40, 42]; the book-length account is Mindlin and Laje [46]. Two scalar control curves, air-sac pressure and labial tension, suffice to reproduce canary song [24], and the bifurcation structure carries the timbre: oscillations born in a Hopf are tonal, those born in a saddle-node-on-a-limit-cycle are pulse-like and harmonically dense, and the spectral content index is a function of tension alone [1, 68]. Practical fitting procedures follow: γ from two (f_0, SCI) points [69], $\chi^2(\gamma)$ over segments [55], and a nearest-neighbour lookup in a precomputed table and not an optimiser [8].

The periphery is a separate and equally well-measured problem. A pure tone in birdsong is a filter effect, not a sinusoidal source: harmonics that appear in a helium–oxygen atmosphere are suppressed in air [5, 51]. The oropharyngeal-esophageal cavity is actively tuned to track the fundamental within a syllable [63], while the trachea is a tube of fixed length that resonates as a stopped quarter-wave pipe at odd harmonics of one body-determined frequency [50, 62]. Beak gape acts as a high shelf on 4–7.5 kHz, not as a movable formant [50], and it covaries with fundamental frequency in song [52, 56]. Finite-element treatments of a real tract exist [21, 36]. Above the periphery, respiration supplies the pressure gesture, with mini-breath and pulsatile regimes separated by syllable rate [31] and a two-state air-sac oscillator now published

[18]. The two syringeal sides are independently controlled [26, 70], they couple through the shared trachea [41], and two-voice complexity can arise from one side alone [77]. Finally, the model needs noise in the right place: matching the fundamental and the filters was not enough to drive song-selective forebrain neurons, and noise on *tension* (at an intermediate magnitude) was required [2]. Real crystallised song carries about 1 % rendition-to-rendition fundamental-frequency variation [72] with slow within-day drift [74], and syringeal muscle can follow modulation to 250 Hz [15, 61]. A different model class, the two-mass mucosa-wave family, is the right one for doves [16].

3.2 Birdsong syntax

The syntax literature establishes that birdsong is not first-order and that the departures are specific. Bengalese finch song has finite-state structure with repeat-count-dependent transitions, and a repeat-count model is the minimal sufficient description [34, 53]; partially observable Markov models improve on hidden Markov models on the same species [44]. Nightingales hold 180–200 song types with package organisation [33] and countersinging rules [32]. On the reference side, the Japanese tit *Parus minor* combines two call elements into a compound whose effect on receivers is lost when the order is reversed [71], Siberian jay alarm calls encode predator behaviour [27], and the black-capped chickadee's *chick-a-dee* call has four note types in strict order with D-count scaling to threat [22]. Berwick et al. [6] is the standard cautionary reading on how far the analogy to human syntax can be pushed, and argues that what is called birdsong syntax is phonological and not semantically compositional, which is the reading this work adopts, since it makes no claim above the level of sequence structure. Beyond birds, sperm-whale codas have been factorised into a combinatorial alphabet using structure-first methods [67], which is the template for doing this without playback. And vocal mimics supply the one form of cross-species ground truth available: lyrebird imitations of grey shrike-thrush song elicit responses from shrike-thrushes as strongly as conspecific song does [14], so a mimic and its model are a pair certified perceptually equivalent by the animals concerned.

3.3 Bioacoustic embeddings and audio foundation models

Unsupervised repertoire mapping is an established method. Sainburg et al. [66] project diverse animal vocal repertoires into two dimensions and quantify latent structure across species; Goffinet et al. [25] learn low-dimensional features that resolve individual and group differences. The interactive descendants of this line. Google Creative Lab's *Bird Sounds*, the TensorFlow Embedding Projector, Apple's Embedding Atlas. Are where the visual language of the "galaxy" comes from.

The supervised and self-supervised model line has moved faster. BirdNET is an open species classifier trained on the same archives, at roughly a hundred times the scale of this corpus, working at 3-second resolution [35]. Perch, and now Perch 2.0, provides a general bioacoustic embedding over roughly 15 000 classes at 5-second resolution [73]. AVES applies self-supervised speech pretraining to animal sounds [29]; BirdMAE is a masked autoencoder for bird audio [59]; NatureLM-audio is an audio-language model for bioacoustics [64]. Benchmarks and standardised datasets have arrived with them [30, 60]. Two findings from that literature are load-bearing here. First, recording-condition invariance is treated as a training-time problem, via domain-invariant contrastive objectives that pair same-class examples across domains [49]; the closest available prior on the recording channel itself comes from *outside* bioacoustics, where convolving training audio with recorded device impulse responses and mixing frequency-band statistics improves generalisation to unseen recording devices in acoustic scene classification [48]. Second, and against the grain of the first, the AVEX study. An empirical sweep of data diversity, scale, architecture and training recipe over 26 datasets. Finds that data diversity dominates: more genuinely different domains per species buys more than the architecture or the loss does [47].

3.4 Methods borrowed from other fields

Four toolkits are used here without modification. Dimensionality reduction and density-based clustering for the map [9, 45]. Continuation-count smoothing for the transition model [37]. Batch-effect correction from genomics and probabilistic linear discriminant analysis from speaker verification, as post-hoc candidates for removing recording-condition structure [38, 57]. Normalised compression distance as a model-free sequence baseline [12], whose underlying normalised information distance is Li et al. [43]. Mel-cepstral distortion as an audible-distance metric [39]. The Menzerath–Altmann law as a falsification test on a unit inventory, which has been applied to gelada and to penguin vocal sequences [19, 28]. Random Fourier features for the browser's kernel ridge [58], a normalised temperature-scaled contrastive loss [11], gradient-reversal adversarial heads [23] and style mixing [76] inside the learned critic. Two methods named as the right next step are simulation-based inference for mechanistic simulators [13] and differentiable digital signal processing [17].

4 The opening this leaves

From the physical-model literature this work takes the oscillator unchanged. The implemented equations are the published normal form term for term, with a sign convention on x that places the limit cycle at positive pressure and tension; that equivalence was verified numerically before the parameter mapping was fixed. It takes the parameter reading unchanged too: α is

air-sac pressure, β is labial tension and sets the fundamental, γ is a pure time-scale. It takes $\gamma = 24\,000$ as the value at which the calibration table is measured, following Perl et al. [55], and the tube formula for the tracheal comb from Nelson et al. [50] and Riede et al. [63]. It takes the placement of noise on tension, not on pressure or on the output directly from Amador and Mindlin [2], and the 140 Hz corner for it from the respiratory/syringeal modulation boundary [61]. It takes the respiratory oscillator and the two-voice coupling kernel as published [18, 41].

Three departures are deliberate and are the origin of most of what follows.

Per syllable, not per bird. The literature fits γ once per bird. Here every syllable gets its own time-scale, chosen so that the syllable's median pitch sits at $\beta \approx 1$. With γ pinned, the oscillator cannot go below roughly 1.4 kHz, and every owl, dove and cuckoo in the inventory lands one to three octaves sharp because the tension floor binds instead. The consequence, discovered later, is that pinning to $\beta \approx 1$ pins every syllable to the *tonal* end of the model and not to its middle, which amputates the spectrally rich tail (Section 7).

At archive scale, from measurements only. The literature fits one bird at a time from retained audio. Here the fit runs over 16 283 types across 2 650 species from descriptors and log-mel patches, with the audio deleted. That forces the objective to be a function of what the store kept, which turns out to be the binding constraint on fidelity.

An inventory and a grammar, not a single song. The unit of the system is a per-species syllable-type inventory with a measured transition table over it, so that the map has named regions and the synthesiser can improvise in a species' own measured sequence statistics, not replaying one fitted phrase.

From the embedding literature this work takes the repertoire-map idea and the finding that recording condition is a first-order confound. It *starts* from the observation that the confound was measured in its own embedding at +0.488 nearest-neighbour lift against species' +0.196, and that this number and not the fit, is what limits every downstream claim.

From the foundation-model literature it takes Perch and BirdNET strictly as external second opinions. Their labels are metadata and are forbidden from becoming typing ground truth: the moment an external inference becomes a type's identity, model error has been laundered into the canonical inventory and the pre-registered-metric discipline stops meaning anything. There is a second reason for that rule which is specific to the resolution mismatch: Perch labels a 5-second window, not a syllable, so background birds can dominate its answer.

5 The system

Figure 2 gives the whole chain. Stages 1–4 run offline in Python; stage 5 is a browser application reading static assets, with no Python process and no model weights.

5.1 Harvest: decode, describe, delete

The harvester fetches from xeno-canto [75] and iNaturalist [10] through one loop that decodes a recording, segments it, describes every syllable, and deletes the file; peak memory is one recording. What is retained per syllable is a 28-dimensional descriptor, a 32×23 log-mel patch (736 numbers, log-magnitude only, no phase), a four-knot mel path, a duration, a pitch contour and provenance fields.

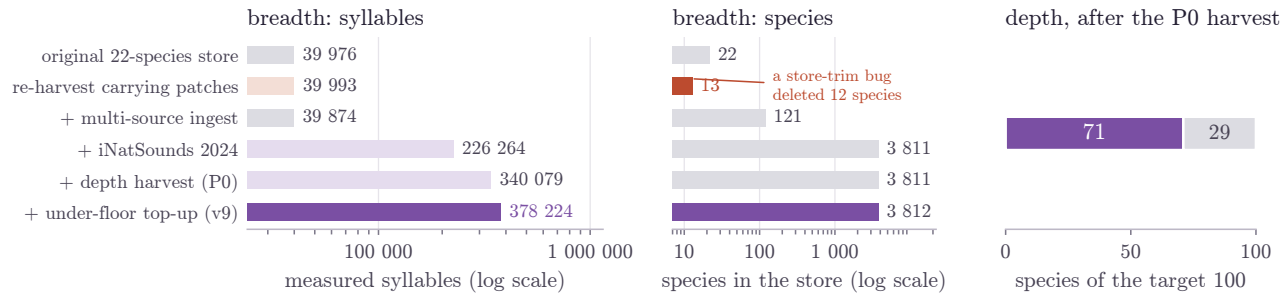
The corpus grew in four documented steps: an original 22-species store of 39 976 syllables; a re-harvest carrying patches; a 121-species multi-source ingest; an iNatSounds pass adding 6 453 recordings, 226 264 syllables and 3 811 species in one run; and a depth harvest adding 139 440 syllables and taking recordings from 7 992 to 10 184. The depth harvest targeted 100 species chosen for documented syntax cases, repertoire depth and existing playback anchors, at up to 30 recordings each. It reached a median of 29 recordings per species on those 100, with 71 above the 20-recording threshold the syntax work needs. The tenth percentile is 5, and that is the archive's ceiling and not the harvest's: xeno-canto holds no more for those birds.

Audio retention was later permitted where the licence allows it, for one measured reason: an external model and any per-syllable re-measurement need the waveform. NoDerivatives is refused, because everything downstream produces derived measurements, and an unrecorded licence is refused, because absence of a licence is not permission. On the production store, 90.3 % of 10 184 recordings pass and all 10 184 carry a recordist. Nothing is shipped to the browser.

5.2 The segmenter band and five dead noise gates

The segmenter is where noise enters, and for most of the project it was listening to the wrong thing. Onsets were thresholded at 14 % of each recording's *full-band* envelope peak, while every description downstream discards everything outside 400 Hz–12 kHz. On a recording with wind, handling noise or traffic under it the two disagree and the segmenter wins: it slices the rumble, and the description then reads whatever tonal haze sits inside those spans. The case that exposed it was found by ear, and 98.7 % of that recording's energy lies below 400 Hz; its 95 stored syllables have a pitch median of 992 Hz against the corpus median of 2 762 Hz. It is not a rare shape. Over a 200-recording sample, the share of energy under 400 Hz runs median 35.5 %, p75 86.8 % and p90 97.5 %, so roughly a quarter of recordings put over nine tenths of their energy outside the band that describes them. Segmenting on the analysis band instead, measured over 120 recordings, gives 22.5 % fewer syllables at a pitch median of 3 179 Hz instead of 2 622 Hz and a signal-to-noise median of 15.9 dB instead of 12.7 dB, improving 81 % of recordings.

That gain replicates, but only as a per-recording statistic, and the corpus has still not



Syllables and species are separate panels because they are separate units; the old figure overprinted one on the other. The second row is drawn in clay because a store-trim bug deleted twelve of its 25 species, so any per-species figure measured on it has chance 7.7 % rather than 4.0 % and is not comparable to the rows around it. Right: 71 of the 100 targeted species reach the 20-recording bar the syntax work needs, at a median of 29; the tenth percentile of 5 is the archive's ceiling rather than the harvest's, because xeno-canto holds no more for those birds.

Figure 1: The corpus grew in breadth and then in depth. Syllable counts and species counts are separate panels because they are separate units. The second row is drawn in clay because a store-trim bug deleted twelve of its 25 species, so any per-species figure measured on it has chance 7.7 % instead of 4.0 % and is not comparable to the rows around it. The right panel is the depth harvest: 71 of the 100 targeted species reach the 20-recording bar, and the tenth percentile of 5 is the archive's ceiling, not the harvest's.

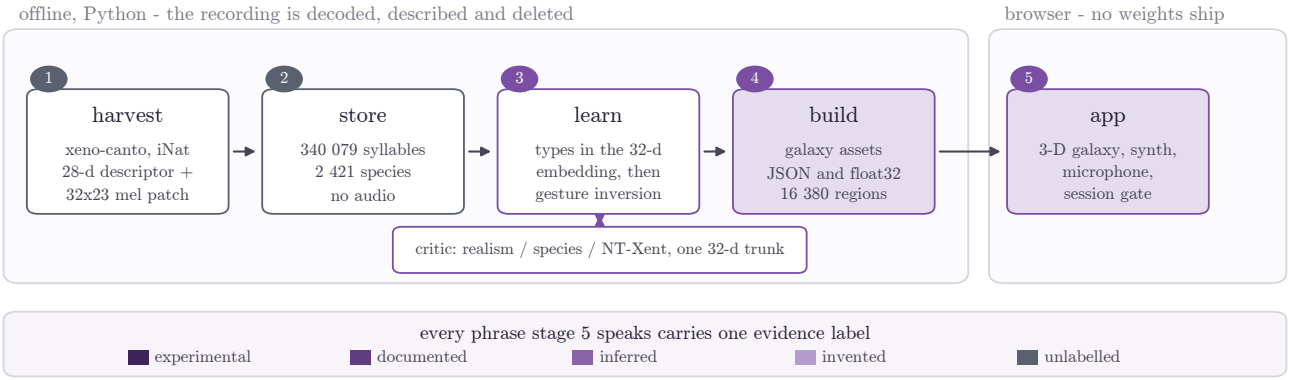


Figure 2: The system end to end. Stages 1–4 are offline and are the subject of Sections 5. Stage 5 reads static JSON and `float32` assets. The evidence labels of Section 1 sit over everything stage 5 displays.

been re-derived. An intermediate re-measurement on 60 recordings took the median over every resulting row instead of per-recording medians and found only 0.77 dB. That figure is withdrawn: the same statistic on 120 recordings from the same cache gives 2.57 dB, because a median over rows lets one recording with hundreds of near-floor syllables outweigh fifty clean ones. The per-recording statistic is stable across both samples at 2.46 and 2.90 dB with 78 and 77 % of recordings improving, which is the 3.2 dB and 81 % above; the absolute pair is sample-specific and we quote the gain rather than the levels. A further check holds the meter fixed. Each arm ordinarily measures its ratio in the band it segments in, so any cross-arm comparison spans two meters; scoring both arms' spans through one common band and one shared noise floor, the gain does not shrink but grows, the full-band arm falling to 2.91 dB against the band-limited arm's 8.95 dB, because full-band spans sit on rumble peaks and are close to the floor in the band that describes them.

What the comparison also shows is the shape of the disagreement between the two segmentations: matched span by span, the rows only the full-band path finds have a median ratio of 4.68 dB and the rows only the band-limited path finds have 2.64 dB, against 13.61 dB for the rows both find. The two paths agree on the strong material and differ almost entirely on marginal spans near the noise floor. Projected over the corpus, re-segmenting costs 22 % of rows and drops species above the fitting floor from 2 648 to about 1 938, against a 6 dB signal-to-noise floor that buys 3.3 dB at 0.77 retention and keeps 224 more species. The two rates are close rather than four to one, and they are not the same kind of number: a floor selects rows on the field it then reports, so its median rises by construction and no span is improved, while the band fix changes which spans exist and does not select on signal-to-noise at all. The corpus is still segmented as harvested, and the choice between breadth and per-recording quality is recorded as open rather than settled.

That was a defect, not a filter, and fixing it did not produce a filter. Five candidate ingest gates were measured against the one recording a human had labelled noise (Figure 3), and four of them place it mid-distribution or in the wrong tail: it reads as *more* tonal than average, and Perch 2.0 identifies a bird in it at the 88th percentile of confidence in the corpus. A threshold that catches it rejects between 43 and 89 %

of everything else. The reason is structural, not unlucky: loudness, tonality, harmonicity, model presence and rhythm each name a property that some real bird has in the extreme. The most metronomic recordings in this corpus are tapaculos, a snipe and a wryneck and not insects, so over 3 581 species no one-dimensional gate has a tail belonging only to noise. Each threshold was also fitted against a single labelled example, which is a description of that example and not a rule. Ingest quality is therefore carried as a *weight* and not as a filter: a signal-to-noise ratio and a clipping flag ride on every observation, and unmeasured is treated as neutral, not as bad.

5.3 The 28-dimensional descriptor

One descriptor per syllable offline, or per rolling window in the browser, computed sample-for-sample identically on both sides; library defaults are not relied upon. Audio is 48 kHz mono, peak-normalised to 0.9, framed at 1024 with hop 512 under a symmetric Hann window. A frame is voiced if its unwindowed RMS exceeds 10^{-4} and 5 % of the clip maximum. Per voiced frame: 13 mel-frequency cepstral coefficients from 26 triangular filters on the power spectrum with energies floored at 10^{-10} and an unnormalised DCT-II; the spectral centroid as $\log_2$ Hz; and spectral flatness. The descriptor is the mean of the 13 coefficients, their population standard deviations, the mean $\log_2$ centroid and the mean flatness, standardised against corpus statistics shipped with the assets.

5.4 The critic: one network, three heads

Everything else in the pipeline is hand-written NUMPY. Two quantities, however, are distances and not physics, and hand-designed distances kept failing. A small convolutional network on the log-mel patch carries three heads off a shared 32-dimensional penultimate layer: a realism logit (real audio against our own resynthesis), a species classifier over the training species, and a contrastive head under a normalised temperature-scaled loss at $\tau = 0.2$ [11]. Species are held out entirely for the realism evaluation, because a network that has seen every species can report identity instead of realism. Augmentations are small time stretches, gain changes and added noise. Never pitch shifts, since pitch is type-diagnostic here.

Forcing all three tasks through one 32-dimensional bottleneck is the point: the embedding has to answer all three at once. The production critic trained on 2 686 species with 895 held out reaches realism AUC 1.000 on held-out species, species accuracy 0.730 over 2 686 classes against chance 0.0004, and novelty ROC 0.846 flagging 32.2 % of unheard species' syllables at a chosen 5 % false-positive rate.

The network never generates. It scores, embeds and flags novelty; what ships is an interpretable parameter vector rendered by the physical model.

5.5 Typing and the fit

Syllable types are k-means clusters in the 32-dimensional embedding, with k chosen per species by silhouette. Replacing the previous hand-weighted seven-number observation vector with the embedding is the largest single accuracy gain in the project (Section 8).

Each type is then fitted to the oscillator. The cluster median supplies a measured pitch contour, duration, timbre class, attack and modulation depths. Coordinate descent (`invert_gestures`) then searches thirteen gesture axes. Including `gammaScale`, `tractQ`, `noise` and `rough`. Over three passes, a few hundred renders per type, moving only on an improvement. Measured contour and duration are excluded from the search deliberately: the pitch tracker reaches them to about 1 %, and a search should not overwrite a measurement.

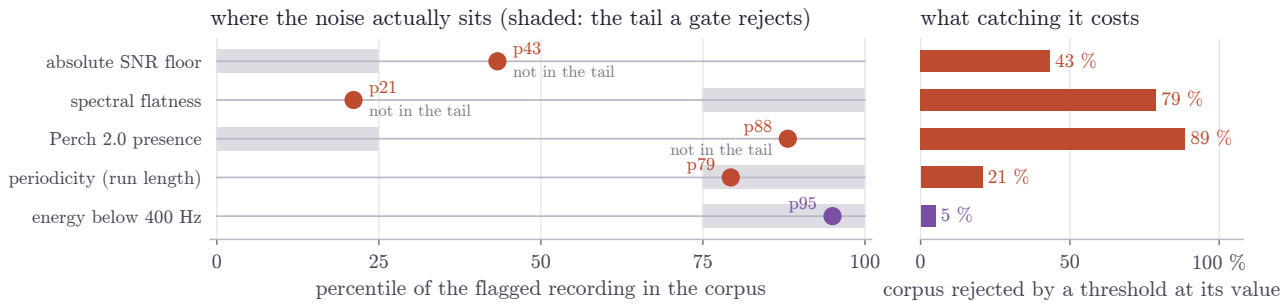
The objective is a cosine between the resynthesis's 28-dimensional descriptor and the type's mean-member reference, in the standardised harvest population frame. That single scalar decides the gesture search, the respiration and two-voice ship gates, and the fricative routing. Section 7 reports that it ranks audible fidelity inversely, and that the respiration gate it decided made the sound measurably worse.

Two per-species fits sit beside the per-type one. The tracheal comb's fundamental f_1 is *measured*, not searched: per-syllable cepstral means are averaged over all of a species' observations, inverted to a 26-band log spectral envelope, detrended, and scanned for the f_1 whose odd-harmonic series best aligns with the residual. Because those syllables span many recordings and a wide pitch range, pitch-dependent structure averages away and only resonances fixed in hertz survive. The long-term-average-spectrum argument applied to a store that kept no audio. The measurement proposes and the resynthesis decides: a species whose recordings show no tracheal colour keeps none. Implied effective tube lengths came out at 1.7–6.6 cm against 3.4–4.5 cm measured in the literature.

5.6 Grammar and songs

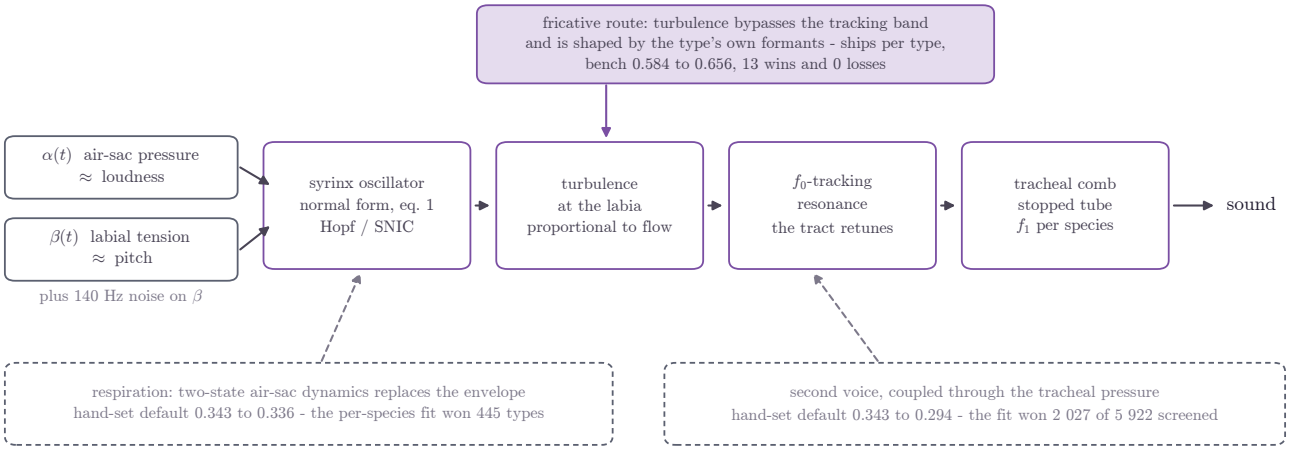
A gap longer than 1.0 s breaks a bout; that constant is shared by the grammar, the chunk miner, the bake-off and the song detector through one named parameter, after a period in which it existed as three separate literals. The shipped grammar is a first-order transition table with Kneser–Ney continuation-count backoff [37], plus 26 897 measured per-transition gaps, so nothing about the rhythm is invented.

Song types are detected, not authored. Each bout's repeat runs are collapsed, the collapsed strings are clustered by normalised edit distance, and a cluster of at least six occurrences becomes a *motif* song type. A second class was added after the funnel was counted: a *repeat* class is one collapsed string heard at least six times in bouts of at least three syllables, matched by equality, because equality is the only similarity a one-token string carries honestly. Both classes ship with a `kind` field so nothing downstream has to in-



The owner identified one ingested recording by ear as noise. Four candidate gates put it mid-distribution or in the wrong tail entirely: it reads as more tonal than average, and Perch 2.0 calls it a bird at the 88th percentile of confidence, so a threshold that catches it takes 43 to 89 per cent of the corpus with it. The one statistic on which it is genuinely extreme is the share of its energy below 400 Hz, and that was not a gate but a bug: the segmenter thresholded on the full band while every description used 400 Hz to 12 kHz, so the syllables had been chosen by rumble that nothing downstream reads. It was fixed rather than gated, because a rumbling recording can still hold a bird. Over 7000 recordings for the three statistics read from stored descriptors; 201 for the two that need the audio.

Figure 3: Five candidate noise gates against the one recording identified by ear as noise. Four place it mid-distribution or in the wrong tail, so catching it costs 43 to 89 % of the corpus. The only statistic on which it is extreme is the share of energy below 400 Hz, which was a segmentation defect, not a gate.



Dashed: implemented and parity-tested, off by default. The published physics is not in doubt; the hand-set default parameter sets lost their pre-registered bake-offs.

Figure 4: The synthesis chain. Two gesture contours drive the oscillator; turbulence, the f_0-tracking resonance and the fixed tracheal comb shape the result. The dashed blocks are implemented, parity-tested and off by default: their published physics is not in doubt, but their hand-set default parameter sets lost their pre-registered bake-offs, and only the per-species fits ship, per type, where they win.

fer which claim is being made. Every configuration is judged against a within-species permutation control that keeps bout lengths and token marginals.

5.7 The map and the live path

The build projects the 32-dimensional embedding to three dimensions, clusters it, and writes plain JSON and `float32` assets: one position per syllable, per-cluster centroids with labels, functions, confidences and citations, and the cross-species relations document. Three clustering modes exist. The default is seeded k-means, where each type contributes a seed and the resulting partition votes a name back by point-count majority. HDBSCAN is the discovery mode, and is the only one that can answer "none of these". Label assignment sends every point that already carries a type to its own type's cluster and fits k-means only over the unlabelled remainder; Section 8 reports what that is and is not worth.

The browser cannot run the critic. No weights ship, which is a hard constraint rather than a packaging detail, so the live path needs a bridge from the 28-dimensional descriptor it can compute to the 32-dimensional space the map lives in. That bridge is a kernel ridge over random Fourier features ($D = 2048$, $\sigma = 4$, $\lambda = 1$), fitted offline and shipped as plain numbers [58]. It is gated on held-out *recordings*, never on held-out syllables, because two syllables of one bird in one take are nearly the same pair twice. A linear ridge came first and was replaced by measurement and not by preference: it cleared the 0.8 gate at 0.8585 on a 19-species critic and fell to 0.7997 once the embedding grew finer.

Live activation is a cosine against every centroid, a softmax at a shipped temperature, and an exponential moving average at 0.82/0.18, which is about 120 ms of smoothing at 60 Hz. A per-point brightness term makes a cluster light from the inside out. The members nearest the live sound brighten first, so an active region

shows its internal structure instead of switching on as one blob.

5.8 Verification

The two renderers must agree sample for sample. A parity harness renders a fixture list through the real AudioWorklet and through the Python renderer and compares them; it stands at 18 of 18 fixtures with the worst divergence at float32 rounding (2.8×10^{-8}). Six divergences had accumulated before it existed, including amplitude modulation applied twice in the browser and a per-integration-step clamp that flattened the high-amplitude regimes. A separate check renders all 44 authored syllables and measures pitch, currently at 0.3 % median error. Forward Euler is validated against Runge–Kutta to within 2 % of f_0, and the calibration table reproduces the reference oscillator measurements to 0.1 % on f_0 and 2 % on peak-to-peak amplitude. The noise generator is part of the contract: xorshift32 with a specified seeding rule, so that "the same turbulence" is a claim two implementations can be tested against.

6 Findings

This section is the answer to one question: what does this corpus actually show. Six things are measured, reproducible, and survive their controls. We state each one first, and put its ceiling in a separate sentence after it, so that a result and its limit are not the same claim.

1. **The learned tokens are units.** A first-order model gains 0.98 bits per token on held-out bouts and loses 0.75 bits on a shuffle of the same tokens.
2. **The embedding holds real per-species geometry.** A species' own types find a held-out type at top-1 0.540 against a chance rate of 0.00007, and the shuffled control sits at chance.
3. **Repeat structure doubles the measurable songs.** Repeat classes take the corpus from 48 to 98 species with a song, and the permutation control does not move.
4. **The published oscillator reproduces the median type** at a descriptor cosine of 0.699 over 16 239 fitted types.
5. **Recording sessions leave a detectable signature**, which is what makes the recordist confound measurable, not suspected.
6. **The corpus has a hard ceiling that is the archive's, not the method's.** Every species below the fitting floor sits at or under 23 observations.

Two conventions apply throughout. Numbers from the 100-species debug subset are marked *debug-scale*, and numbers from twelve-species or ten-type samples are marked as such, because this project has twice measured an effect at small scale that did not survive replication, in both directions. Where a result was measured before the 2026-08-03 top-up, it carries that store's totals of 340 079 observations and 2 421 fitted species. Those measurements were not re-run, and restating them under the current totals would be a claim nobody made.

6.1 The tokens behave as units

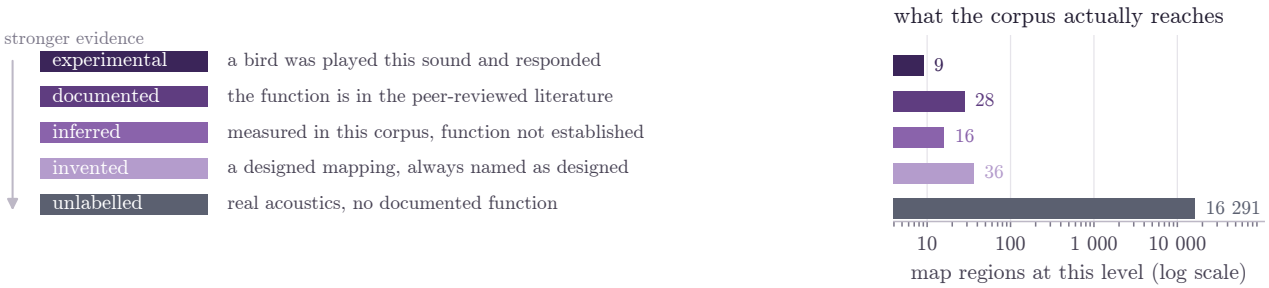
A first-order transition model over the learned syllable-type inventory buys **0.98 bits per token** on held-out bouts, a cross-entropy of 3.284 down to 2.301. On a within-species shuffle of the same tokens, with bout lengths and marginals preserved, the same model *costs* 0.75 bits, 2.879 up to 3.634. The corpus is 34 696 bouts, 327 164 tokens and 2 421 species; the estimator is unbiased by construction, which the plug-in estimator this replaced was not.

The logic is what makes it strong. An unbiased estimator paying for parameters it cannot use is the clean form of the claim that the tokens are units: the model can only win if the transitions are real, and it wins by a bit per token. This holds a load-bearing assumption that everything above the syllable rests on. The syllable inventory is produced by k-means in a learned embedding, which is a clustering; the question of whether it is also a *unit system* is not answerable by comparing two models, and it is answered here in the affirmative. A model-free estimate on the same corpus agrees: conditional entropy of the next token given one previous is 0.941 bits against the shuffle's 3.343.

Two artefacts belong with the number. The control's unigram cross-entropy sits *below* the real corpus's (2.879 against 3.284), because shuffling spreads each species' tokens evenly across its bouts so that train and test marginals nearly coincide, whereas real bouts are clumped (a bout tends to repeat one type) so a held-out real bout can be mostly a type the train half barely saw. That clumping is itself structure, and it is why the gate reads within-corpus deltas and not the two absolute curves against each other. And the same behaviour is pinned in a self-test: a planted second-order process must be found, a first-order process must not show depth, and the shuffle of the planted process must not gain from a longer context.

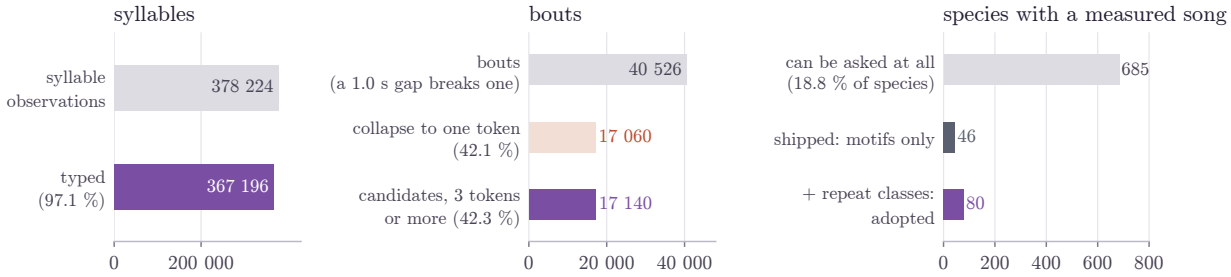
6.2 Repeat classes roughly double the corpus's songs, with the control unmoved

The song funnel was counted stage by stage on the production store, and the loss is accounting, not machinery. Of 340 079 syllable observations, 327 164 (96.2 %) are typed. Those form 34 696 bouts. **19 738 of those bouts (56.9 %) collapse to a single token**, and 6 940 of them are three syllables or more. A bird repeating one sound, which is a repertoire fact and not a measurement error. Only 10 550 bouts (30.4 %) survive as candidates, 366 species have enough of them to ask the question, and 48 species got a song type.



Every phrase the application speaks carries one of these five labels; nothing renders without one, and the weakest level is protected, because detecting that a sound is new is not knowing what it means. The distribution is the point: 16 291 of the production build's 16 380 regions are unlabelled, and the interface is required to keep saying so.

Figure 5: The five evidence labels, the requirement for each, and the corpus counts at the time of writing. Every phrase the application speaks carries one of them; nothing renders without one.



The funnel is drawn in the three units it actually changes through, not on one axis. The loss is accounting and not machinery: 3 215 of the collapsing bouts hold three syllables or more, a bird repeating one sound, which is a repertoire fact and not a measurement error. Repeat classes - one collapsed string heard at least six times, matched by equality - make 91 of the 187 song types, beside 96 motifs, and the song layer places 2 299 bouts. Eighty of 2 650 fitted species is 3.0 %. v9 store, every panel recomputed together by analysis/song_funnel_stages.py; the within-species shuffle control was measured pre-v9 at 0.170 % against a 0.2 % bar and has not been re-run.

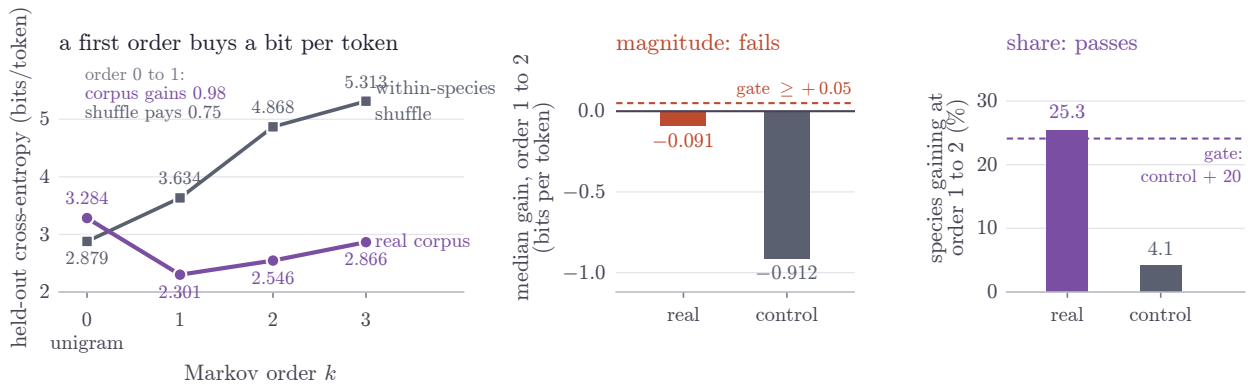
Figure 6: The song funnel on the v9 store, drawn in the three units it actually changes through. Of 378 224 syllable observations, 367 196 (97.1 %) are typed and form 40 526 bouts. Of those, 42.1 % collapse to a single token and 42.3 % survive as candidates. 685 species hold enough candidates to ask the question, and 46 got a motif song type; adding a repeat class (one collapsed string heard at least six times, matched by equality) takes that to 80. **The funnel widened at the bout stage and narrowed at the song stage.** Against the pre-v9 store the bouts rose from 34 696 and the candidates from 10 550, while species with a song fell from 98 and placed bouts from 3 741 to 2 299 — a finer typing splits one physical repeat across several type ids, so fewer bouts clear the six-rendition floor. All three panels were recomputed together from one store by `analysis/song_funnel_stages.py`, which imports the build's own bout rule.

The collapse is correct for a motif distance. A trill sung twice at two speeds must not sit twenty edits apart, and wrong as a filter. Until this was fixed, *a bird whose song is a trill could not have a song type at any recurrence bar.* Adding repeat classes takes the corpus from **48 species with a measured song to 98**, song types from 96 to 356 (96 motifs, 260 repeats), and placed bouts from 1 052 to 3 741, with the shuffle control **unchanged to the bout at 0.170 %** against a 0.2 % bar.

The funnel was recomputed on the v9 store and the figure now shows those numbers, which are not these. The counts above are what the adoption decision was made on, so they stay as the record of it. On v9 the funnel is wider early and narrower late: 40 526 bouts and 17 140 candidates against 34 696 and 10 550, but 80 species with a song against 98, and 2 299 placed bouts against 3 741. The cause is the one `song_funnel.py` exists to catch — a finer typing splits one physical repeat across several type ids, so a bout that collapsed to one token six times now collapses to two tokens four and two times, and neither clears the six-rendition floor. Repeat classes still roughly double the songs at v9, 46 species to 80 and 96 song types to 187, so the adoption is not overturned. But "double" is a pre-v9 measurement and the v9 factor is 1.7.

The control is the reason to believe it. Shuffling a species' tokens destroys runs, so the control finds no repeat classes *at all*. A statistic whose null yields zero, on a corpus where the real version yields 260 types, is about as sharp as evidence gets from a permutation design. And the comparison is a comparison: the baseline arm reproduces the shipped corpus exactly, at 48 species, 96 types and 1 052 placed bouts.

The 0.2 % bar those arms are judged against was itself set on a control sweep, not chosen, and the sweep is worth quoting because it shows the shape of the trade-off. On an earlier 35-species store, lowering the minimum cluster size from 6 to 4 took the real yield from 50 song types to 113 and the control's from 1 spurious type (6 of 3 386 bouts, 0.2 %) to 17 (103 of 3 386, 3.0 %). Raising it to 8 gained the control nothing and lost half the real types. Six is the knee, and the residual 0.2 % is reported and not tuned to zero, because the price of zero was half the signal. Two further constants exist for measured reasons: bouts collapsing



Lower is better on the left. Per species, an 80/20 split by bout, a k-order backoff model fitted on the train half, every held-out token scored. An unbiased estimator paying for parameters it cannot use is the clean form of the claim that the tokens are units: the model can only win if the transitions are real. A second order is not supported - the median species loses 0.091 bits per token at three independent splits - while a quarter of species gain something, six times the shuffle's rate. 217 of 2 421 species clear the 20-bout bar at which the question can be asked. Production scale.

Figure 7: Held-out order scaling. A first-order transition model over the learned syllable-type inventory buys 0.98 bits per token on bouts it has not seen, while on a within-species shuffle of the same tokens, with bout lengths and marginals preserved, the same model *costs* 0.75 bits. The estimator is unbiased by construction, so a model can only win if the transitions are real. A second order is not supported: the median species loses 0.091 bits per token, which fails the magnitude criterion, while a quarter of species gain something, which passes the share criterion. Production scale.

below three tokens are set aside because chance two-token strings sit at normalised edit distance 0.5, and a cluster whose members' weighted mean distance to their medoid exceeds 0.5 is demoted to noise, because five random six-token strings were once reported as one song type.

The ceiling is equally clear. Ninety-eight of 2 421 species is 4.0 %. Two attempts to raise it by loosening a threshold both bought chance strings faster than songs (Table 7). The remaining route is depth, and depth has an archive ceiling.

6.3 No second order, on the 217 species that can test for one

The median species *loses* 0.091 bits per token going from order 1 to order 2 on bouts it has not seen, at three independent splits ($-0.091/-0.116/-0.142$). For the median species a second-order model is worse out of sample: the corpus cannot estimate the contexts it asks for.

The share criterion tells the other half. 25.3 % of species gain something at order $1 \to 2$, against the shuffle's 4.1 %. A six-fold rate that clears the pre-registered 20-point margin. So, in one sentence: *there is no second-order structure this corpus can reach, and a quarter of species hint at one.* The hint is not nothing; it is also not what the gate asked for, and the gate was written first.

What this settles is the diagnosis. The depth question is **data-limited, not method-limited**: 217 of 2 421 species have enough bouts to ask it at all. Given that the literature is unambiguous that birdsong is not first-order [34, 44], the 25 % that hint at a second order are an argument for depth on more species, not for a cleverer sequence model. The post-depth-harvest bake-off supports the same reading from the other side: with qualifying species at 104, not 30, a hidden-state model beats even a smoothing-repaired first-order baseline by about +0.37 nats per token, where at debug depth its entire advantage was smoothing shape.

6.4 The map's own confusion bounds the interaction

With every type granted its own cluster, so that no type can fail to be retrieved for want of one, **type retrieval is 0.602**. For 40 % of points the nearest centroid by cosine, which is the exact lookup the browser performs on live audio, belongs to *another type*. This is debug-scale, on a fixed 214 154-point matrix over 2 143 types.

This is the number to quote when asked whether singing at the system lights the right region. It is not the clustering's failure and it is not the projection's: it is the embedding's own confusion between types, and it was invisible under k-means because a type with no cluster cannot be retrieved at all. Raising it is a representation problem, which is where Section 9's external comparison points.

Two related numbers complete the picture and are better than the headline suggests. On the field points alone. The population the question is about, since a recorded syllable must light its own region. Retrieval is 0.7604 under label assignment and 0.7176 under the shipped k-means (Table 4). And the browser's bridge, which is a second and independent source of error, loses 38.6 % of region agreement of which 76.4 % is boundary near-ties: tie-adjusted agreement is 0.91, and species-level agreement, which is the single-label claim we make, is 0.781 against chance 0.033.

6.5 Pathological recording sessions are detectable from the browser's side

Eight of 45 held-out recordings carry 22.9 % of the map's total flip loss, all eight from xeno-canto and none from iNaturalist, with one session alone carrying 9.3 %. They separate cleanly on a statistic the browser can compute from the 28-dimensional side alone: the mean 5-nearest-neighbour distance of a session's standardised descriptors to the fit-side cloud separates them at **AUC 0.953**, and a shipped cloud of 256 random anchors holds AUC $\geq$ 0.92 over draws at a cost of 28 kB. The pre-registered adoption gate for using it as a *lever* (raising the remaining rows' agreement by 5 points) failed at +3.5 points raw, so the pre-registered fallback fired and it ships as an *honesty* gate: the interface says the session is off everything the map knows instead of lighting regions. At the shipped operating point it flags 8 of 45 sessions, catches 5 of the 8 bad ones including the three worst loss-carriers, and produces 3 clean false flags all within 0.02 of the bad line.

Two calibration facts belong with it. A threshold calibrated on fit sessions does not transfer to unseen sessions even when the session's own rows are excluded, because sibling sessions of the same recordists deflate every fit session's distance. The 95th-percentile fit-side threshold flags 29 % of held-out sessions. The shipped threshold is calibrated cross-wise, each fit session judged against the other seeded half. And the mirror-image action fails: pruning the sessions this signal flags from the *fit* costs −0.0147 and −0.0062 of evaluation cosine on two stores against a 0.005 allowance. The pathological sessions are manifold coverage, and removing them hurts the next unseen pathological session.

6.6 The oscillator reproduces the median syllable type

The audit's conclusion holds and is unusually well supported for a negative: *nothing the project has measured points at the oscillator.* The ODE is the published normal form term for term, agreeing with Runge–Kutta to 2 % of f_0, reproducing the reference measurements to 0.1 % on f_0 and 2 % on amplitude, holding at 18 of 18 two-sided parity fixtures, and collapsing into period doubling exactly where the coupling literature says the weak-coupling approximation dies. Every recorded deviation is in parameterisation, in the periphery, or in the objective.

The concrete state of the fit, at production scale: median resynthesis cosine **0.699** over 2 650 species and 16 283 types. The path to that number is worth stating because it is not monotone and the reasons differ: the 12 724-type learn scored 0.650, the per-species respiration and two-voice fit passes took it to 0.6623 across 12 680 scored types, and it then fell to 0.648 when the corpus deepened, which is the expected direction. The added species are harder. On the original 22-species store the equivalent figure reached 0.775 after the naturalness fields were added, from 0.533 before, with the share of types above 0.7 going from 26.2 % to 66 % and the share below 0.5 from 44.2 % to 11 %.

The single most stubborn number in the project is cross-species placement. On the original 22-species store a resynthesis lands nearest its *own* species 23.1 % of the time against chance 4.5 %. Four hypotheses for the remaining 77 % were each tested and each refuted, and they are worth listing because each cost real work: the envelopes are not collapsed (we express 103 % of the real between-species spread and 106 % of the spectral detail. They vary along different axes); there is a systematic spectral tilt and correcting it *exactly* drops placement to 20.2 %; the harmonic structure is not missing, because real syllables resolve to a mean of 2.2 partials and feeding measured partial amplitudes to the fit moved the cosine 0.776 to 0.768; and a time-resolved objective improves its own metric by 36 % while dropping 28-dimensional placement from 61.7 % to 43.2 %.

That last refutation is the informative one, and its mechanism was found afterwards. The two objectives conflicted because the f_0-tracking tract locks spectral shape to pitch, so the only lever the fit had for moving a spectral trajectory was bending the pitch contour, and the pitch contour is strongly species-diagnostic. Under a fixed waveguide the conflict disappears and all three metrics improve together. That the waveguide then failed to generalise across 22 species, while winning under learned typing on a different store, is recorded as unresolved.

One further reading of the placement metric is required, and it is the reason it is not the paper's headline. *Placement degrades in absolute terms as the corpus broadens, while improving relative to chance.* On the ingest that took the store from 36 to 121 species, own-species placement fell from 24.6 % to 11.5 %, while chance fell from 2.8 % to 0.9 %, so the ratio to chance rose from 8.8-fold to 12.8-fold. The pre-registered target for that gate was an absolute 35 %, and it **failed**. Own-type placement held at 62.6 % and the median cosine held, barely, at 0.742 against a 0.74 bar, with the reason named at the time: 85 of the newly added species carried about three recordings each, which fits thin types. Type coverage in the same build fell from 94.6 % to 54.1 % and cluster purity from 56.7 % to 30.3 %, and both are harvest depth instead of machinery: 651 orphan types belong mostly to those three-recording species, which cannot meet the map's minimum cluster size. Any absolute placement or purity figure in this paper is therefore only interpretable against its own species count.

6.7 Extensions ship only where a fit wins them

The pattern is consistent enough to state as a result about method. **Hand-set physical defaults lose bake-offs, however correct the published physics.** The respiratory oscillator's hand-set default set moved median cosine from 0.343 to 0.336, with losses concentrated in sub-50 ms notes whose onsets the physically correct 20–40 ms attack eats. The two-voice

coupling's hand-set default moved it from 0.343 to 0.294, while the two worst-fitting low-pitched species *improved*, which is why it was kept and not deleted. Both are parked off by default.

The per-species fit passes are the version that wins, and they win narrowly and specifically: respiration took 445 of 12 680 scored types and the second voice 2 027 of 5 922 screened types, with the ship gate holding everywhere. The tracheal comb won 10 of 22 species on the original corpus. The fricative route. Turbulence routed around the f_0-tracking band into the syllable's own fitted formants, because a fricative is a noise burst filtered by the tract and not a tone with noise added. Moved the bench from 0.584 to 0.656 with 13 wins and 0 losses, and ships per type. A one-tone re-fit of the authored calls against the same bird's learned types shipped a measured noise/roughness texture to 14 of 15, median gain +0.148, with the caveat that the median cosine afterwards is still 0.270.

Every one of those gates was read on the descriptor cosine, and Section 7 says that cosine ranks audible fidelity inversely. Two of those gates have since been re-judged there and they disagreed: respiration lost, at 33.1 % of 175 types improved and −4.18 dB of median distortion, while the two-voice fit held, at 55.6 % of 1 737 and +1.13 dB. The other two still owe that rejudgment and both need a refit to get it.

6.8 Novelty detection improves with corpus breadth, but its threshold does not transfer

Novelty detection is a distance question in the embedding, and it is evaluated by holding entire species out of training and asking whether their syllables are flagged. The held-out figure improved monotonically as the corpus broadened, and every step changed the test as well as the model, so the four numbers are a trend, not a paired comparison. On the 9 000-syllable store with six species held out: ROC AUC 0.596 at 7.4 % true positives at the 5 % false-positive operating point. On the 47 000-syllable retrain with five held out: 0.672 and 14.8 %. On the 121-species store, evaluated against 25 892 syllables of 105 genuinely non-training species, not a handful of chosen holdouts: 0.786 and 22.3 %. On the production critic, with 895 whole species held out. The hardest test run: **0.846 and 32.2 %**. The novel side of the test got larger and stricter at each step, and the doc that reports each pair says so.

An earlier and much higher figure, ROC 0.825 on the 9 000-syllable store, is a different measurement. It scored held-back syllables of training species against held-out species' syllables, and the 0.596 above superseded it in place. Both are on the record. The low-false-positive end of the curve is the weak end in every version.

The per-species pattern refutes the obvious explanation. Detection rate is not about taxonomy. On the 9 000-syllable store the idea that congeners are harder fits five of six held-out species. The sixth contradicts it, because a wren with no relative in the training set is the *worst* detected of the six at 8.2 %. The ordering differs on the 47 000-syllable store, which is itself the point: the embedding measures acoustic neighbourhood and not relatedness. A loud broadband trill sits near other species' trills whatever the family tree. That is arguably the correct behaviour for the question being asked, and it means "hold out a species" is a harder test than "hear a genuinely new sound".

The operational finding is a warning. *A novelty threshold holds only for the corpus that calibrated it.* A threshold calibrated for 5 % false positives on the 9 000-syllable store flagged 29.8 % of the 47 000-syllable store, including 29 % of *training* species' syllables. Recalibration without retraining returns the store-wide rate to 5.6 %. The shipped threshold has moved through 0.1048, 0.1558, 0.1912, 0.1737 and 0.1323 across five harvest epochs. Thresholds do not transfer, and treating one as a constant is how a map starts confidently reporting novelty about sounds it already knows.

6.9 One example of ground truth from outside, at n = 31

Vocal mimics are the only source of labelled cross-species acoustic similarity available to a corpus like this one, because a lyrebird singing a grey shrike-thrush produces syllables whose acoustics are shrike-thrush and whose label is lyrebird, and the accuracy of the imitation has been certified by the receiving birds, not by us [14].

The pilot is a well-formed test on an empty corpus, and is reported as such. The store holds **31** lyrebird syllables. Against the 12 393 type centroids of the production fit, those 31 match **13 distinct non-lyrebird species**, where random syllables of the corpus match 29.9 (sd 1.0). Concentrated, in the direction mimicry predicts. Two of the top matches are Australian species that co-occur with lyrebirds and are plausible models; the rest are Neotropical and African and cannot be mimicry. The accompanying figure of 83.9 % for "nearest centroid is its own type" is circular by construction, since those syllables are inside those centroids, and is not evidence.

At $n = 31$ this licenses nothing beyond a direction. The action it names is a targeted harvest, not a method: xeno-canto's lyrebird recordings frequently name the mimicked species in the remarks, so such a harvest returns *labelled* mimic-to-model pairs, and other mimics extend the same set. The reason to pursue it is that it would also supply a fidelity yardstick for the objective problem of Section 7 that owes nothing to our synthesiser and nothing to our cosine.

6.10 Questions this corpus cannot answer

Stated as a list, because each item is a boundary and not a shortcoming to be worked around.

Function, for almost everything. Playback is the only proof, there is none here, and every recording is

one bird with no interaction and no receiver. The interface correctly reports "real acoustics, no documented function" for the overwhelming majority of the corpus: at the production build, 13 919 of 14 003 regions are *unlabelled* against 27 *documented*, 10 *experimental*, 16 *inferred* and 31 *invented*. The store holds 23 697 syllables the novelty detector has flagged, of which 841 are reported as *unlabelable* even by an external model, because their xeno-canto audio was deleted by design. A cost of the retention policy, stated, not absorbed.

Species replication on retained audio. The cache holds at most two recordings per species, so any test needing several recordings of one species requires a refetch.

Geographic priors. BirdNET's occurrence metamodel takes latitude, longitude and week, and the store keeps date and country but *no coordinates*. Both archives return them; the harvester never recorded them. That is a metadata-only re-fetch and it is the gate on the only part of the external comparison still unanswered.

Individual identity and dialect. The corpus now has the depth in principle, 10 to 16 individuals per species is literature-proven separable, and the recordist confound of Section 7 makes any such claim uninterpretable until it is addressed in the data.

7 The contribution and its five limits

7.1 The contribution, stated plainly

Four things here are not standard practice in this area.

A physical inventory at archive scale. 16 283 syllable types across 2 650 species, each one an interpretable parameter vector for a published oscillator, not a latent code. To our knowledge no comparable inventory exists; the physical-model literature fits individual birds.

A corpus that is a description and not a copy. The store holds no audio and cannot be inverted to recognisable audio, which changes what can be shared and removes the licensing question from the analysis entirely.

Evidence labelling as an enforced invariant. Five levels, applied to every rendered phrase and every named region, with the "unlabelled" level protected: detecting that a sound is new is not knowing what it means, and the interface is required to keep saying so.

Pre-registration as the working method. Fifty-eight gates, written before their runs, with the refusals kept in place. This is why the rest of this paper can be specific about what the data refuses.

7.2 Limit 1. The fit objective cannot hear

The pipeline's universal fit objective is a 28-dimensional descriptor cosine. We measured it against two independent audible-distance metrics over 200 fitted types: mel-cepstral distortion [39] and multi-resolution log-spectral distance. The cosine carries almost no information about either. Against MCD it reaches Pearson +0.106 and Spearman +0.100. Against STFT error it reaches Pearson +0.102 and Spearman +0.111. Neither is distinguishable from zero at that sample, and both point estimates lean the wrong way. The instrument itself is sound. The recomputed cosine reproduces the stored one at Pearson 0.921, and the two distortion metrics agree with each other at 0.925.

At six times the sample the wrong sign becomes significant, and the claim gets worse. Rebenched on the production store over **1 200** fitted types — a different population from the 200 above, so this is a second measurement and not a re-run — the cosine's correlation with audible distortion is no longer indistinguishable from zero. It is *positively* correlated with distortion on both independent metrics:

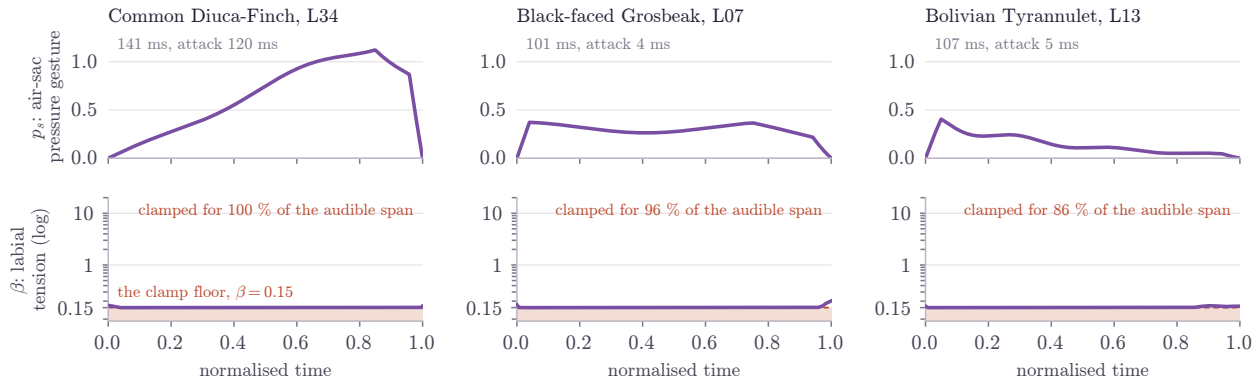
$n = 1\,200$	Pearson	Spearman
cosine vs MCD	+0.098	+0.200
cosine vs STFT	+0.091	+0.186
mel target vs MCD	−0.346	**−0.456**
mel target vs STFT	−0.312	−0.440

Every one of those eight is significant. The two that matter most: the cosine's Spearman against MCD at $p = 3 \times 10^{-12}$, and the mel target's at $p = 1 \times 10^{-62}$.

Higher cosine goes with *more* audible distortion, on two metrics that were built independently and agree with each other at 0.925. At $n = 200$ that read as a point estimate leaning the wrong way. At $n = 1\,200$ it is $p = 3\text{e}{-12}$. So the objective is not merely uninformative about fidelity, which is what this section claimed before: it is mildly anti-correlated with it, and the effect survives a sixfold increase in sample on a different store. The mel target's ranking advantage survives the same increase — Spearman −0.456 against the cosine's +0.200, opposite signs, both significant.

One caveat on the population. Of 2 396 types considered, 1 196 were skipped: 753 have no cached audio and 443 are too short to bench. The benched set therefore over-represents types whose members survived the audio cache, and nothing here corrects for that.

That caveat has since been removed, and the result replicates without it. Caching the store's own xeno-canto audio takes the missing-audio skip to **zero**: a rerun benches 1 200 types with nothing skipped for want of audio and 347 skipped only as too short, on references that are 970 lossy and 230 lossless. Because per-file band limiting makes rows from archives an octave apart incomparable, every row in that rerun is instead held to one common 11 025 Hz band. On that population the cosine's Spearman against mel-cepstral distortion is +0.233 at $p = 3 \times 10^{-16}$ against the +0.200 above, and the mel target's is −0.385 at $p = 1 \times 10^{-43}$ against −0.456. Both signs and both magnitudes hold, so the wrong-way correlation is a property of the objective and not of which



The two gestures the oscillator is driven by, for the same three types as the acoustic examples, read straight out of the shipped fit through the repository's own render planner and its measured tension table. This is what interpretable parameters buys: every sound in the corpus is two named physical contours and thirteen searched gesture axes, not a latent code. Pressure and tension are separate panels because they are separate units. The clamp annotation is measured per type, weighted by where the envelope is actually open; across the production learn the clamp bites in 6 617 of 12 724 types.

Figure 8: The fitted gestures behind the three acoustic examples. Every syllable in the corpus is two named physical contours (air-sac pressure and labial tension), not a latent code, and this is the whole of the interpretability claim. The dashed line is the tension clamp of the parameterisation; a contour that rides it is at a limit of the parameterisation and not at a fitted value, which happens in 6 617 of 12 724 types across the production learn.

archive supplied the audio or of how each row was banded.

Pooling those references is licensed by measurement rather than assumed. At the common band the lossless-backed rows still read a median distortion of 71.49 dB against the lossy-backed 95.23 (Mann–Whitney $p = 3 \times 10^{-6}$), which would ordinarily forbid averaging them. That residual is population: scoring *one* render of the same type against a lossless and a lossy reference, across 273 types that appear in both archives, moves the median by +0.72 dB with the lossy side worse on 53.1 % of types ($p = 0.24$). Holding the type fixed, the archive does not matter; the arm gap records which birds each archive happens to hold.

The mechanism is legible. High-cosine types are clean whistles. Their renders concentrate energy in one band while the field audio carries broadband noise everywhere else, so the spectral distance is large. Noisy buzzes score a low cosine, but they carry broadband energy on both sides, so their distance is small.

The obvious replacement was tried and does not work as an optimiser. A time-resolved mel-path target correlates with MCD in the right direction and at usable strength (Pearson −0.305, Spearman −0.420), which is what a fidelity ruler should do. Blended into the objective at weight 0.5 it wins 31 % of types on MCD against a pre-registered 60 % bar, with a median ΔMCD of exactly 0.0 dB and a maximum cosine loss of 0.120. A ten-type probe had won 7 of 10; the 80-type gate run settled it.

The project therefore measures spectrally and optimises descriptor-wise. Every cosine-gated adoption inherits that caveat, and we record it here. The respiration and two-voice fit winners, the fricative routing and the one-tone texture re-fit were all decided by a proxy that ranks audible fit at type level *inversely*. Each of them also improved the descriptor match, and "improved the proxy" is a weaker claim than it reads.

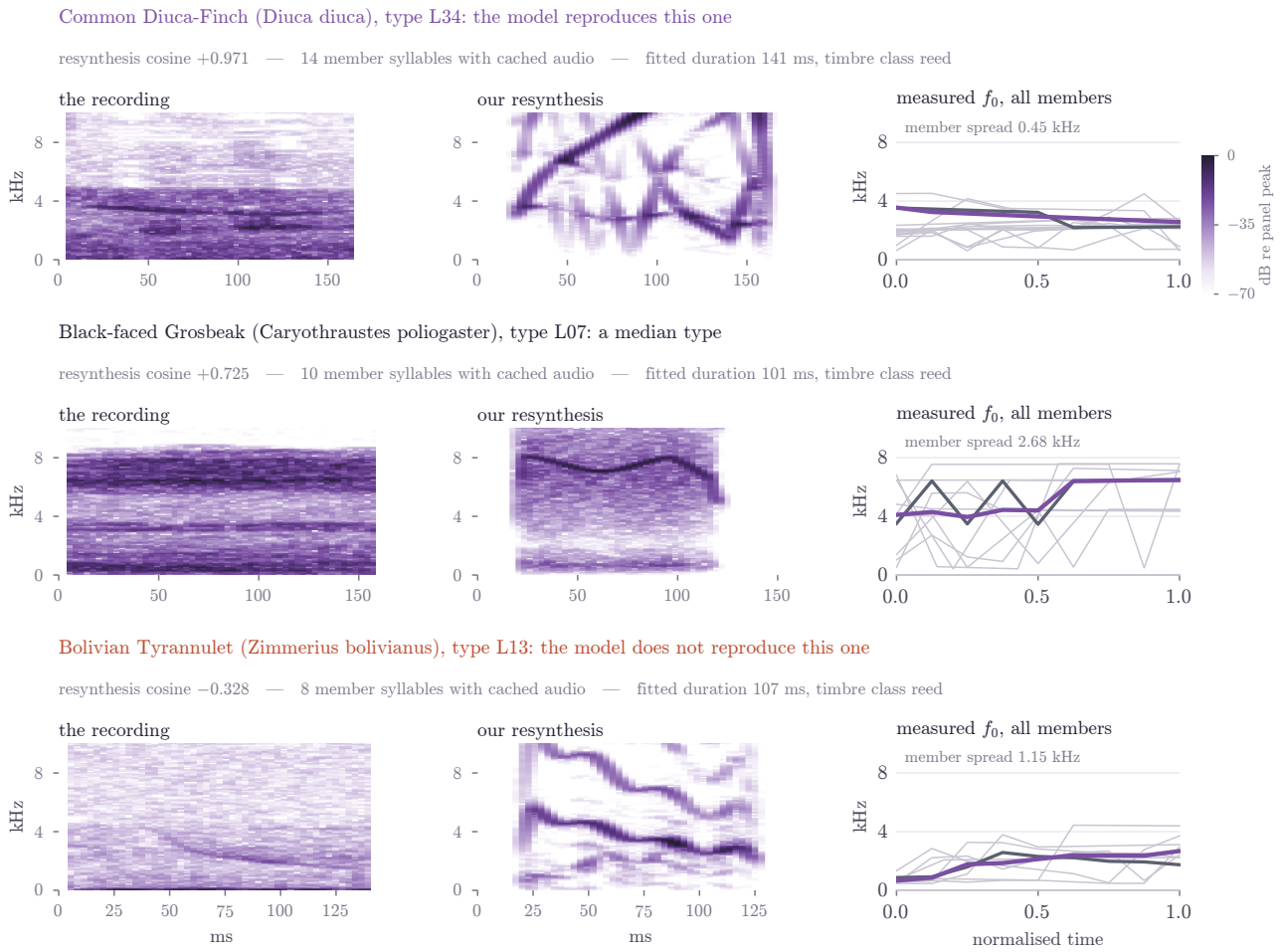
Two of those four adoptions have now been re-judged on the bench. They disagree. The respiratory block was the first, and it made the sound worse. `respfit` ships it on a type only where it beats that type's cosine by 0.02, and in the shipped inventory the field is present on *exactly* the 204 types where it won and on none of the others. So stripping the field reconstructs the pre-adoption syllable and leaves everything else identical. Both arms were rendered and benched. Of the 175 types that benched in both:

median Δcosine (the gate's own number)	**+0.034**
MCD win share	**33.1** % (58/175)
median ΔMCD	**−4.18** dB
mean ΔMCD	−18.40 dB
median ΔSTFT error	−2.34 dB

Positive ΔMCD would mean the adoption reduced audible distortion. It is negative on the median and on the mean, negative on the second metric too, and it helped only one type in three. **An adoption that gained a third of a cosine point degraded the audible fit by about 4 dB.** This was pre-registered before the run — given the cosine's positive correlation with distortion above, the prediction was a win share below 50 % with a median ΔMCD at or below zero, and both held.

Two things this does not say. It does not say the respiratory model is wrong: the physics is from Fainstein et al. [18] and this measures a gate, not a mechanism. And it does not generalise, which the second re-judgment then demonstrated.

The two-voice adoption was re-judged next, on twelve times the population, and it held. An earlier version of this section reported that the two-voice block could not be stripped, because `two` is present on winners and non-winners alike. That was the wrong field. `two` is a boolean axis of the gesture descent, searched for every type; the *fitted* second voice



Each spectrogram: 22.05 kHz cached audio, 256-sample Hann window, 64-sample hop, peak-normalised, 70 dB range. Both sides of a row carry identical settings, identical axes and one colour scale. The resynthesis is rendered by the repository's own renderer at 48 kHz and resampled to the cache's rate; 20 ms of context is kept on the recording and matched with silence on the resynthesis, so the flat band at each panel's edge is the field recording's own noise floor. Third panel, grey: the measured pitch contour of every cached member of the type; slate: the member plotted to its left; violet: the contour the fit ships. The cosine is the objective the fit was run under; docs/19 measured its correlation with audible distortion at about +0.1, so it ranks the fit's own criterion and not audible fidelity. Production store, and the audio cache holds iNaturalist recordings only.

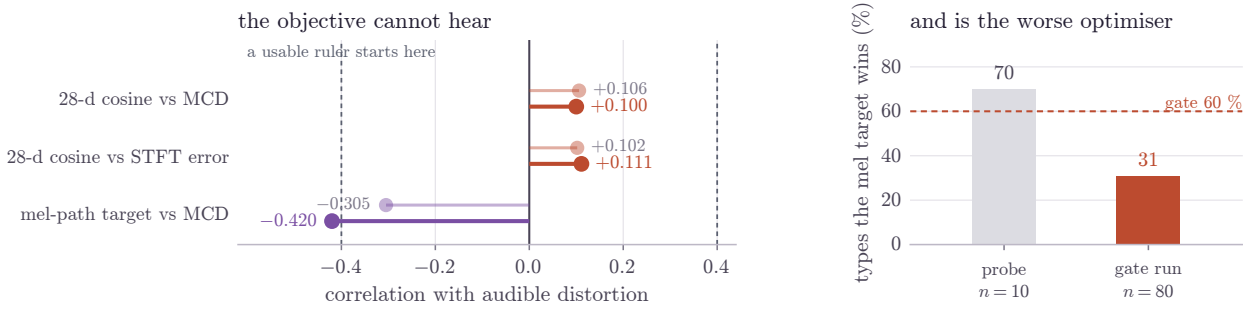
Figure 9: What the physical model reproduces, and what it misses. Three fitted types, taken from the top, the middle and the bottom of the resynthesis-cosine ordering over the types whose members have cached audio; the corpus median is 0.699 over 16 239 types, so the middle row is a median type, not a chosen success. Top: a clean descending whistle, which the oscillator and its tracking tract reproduce closely. Middle: the median case. The pitch is right, and the pulse train that carries the real syllable's timbre is replaced by a steady tone. Bottom: a failure, and the third panel says why. The type's member contours disagree by kilohertz, so the cluster median the fit starts from describes no member well. The cosine is the objective the fit optimised and ranks audible fidelity *inversely* (Section 7).

is `voices`, and `respfit` writes it exactly the way it writes the respiratory block — onto a copy of the syllable, kept only where the gain clears the margin. So it is additively reversible on the same argument. That is checked against the run's own report rather than asserted: 2 144 shipped syllables carry `voices`, the report records 2 144 two-voice winners, and the sets agree with no member on either side alone and no block differing. Of the 2 144, 1 737 benched in both arms, and the smallest Δcosine among them is +0.020 — the gate's own margin, which is what says the arms are paired correctly.

median Δcosine (the gate's own number)	**+0.071**
MCD win share	**55.6** % (966/1 737)
median ΔMCD	**+1.13** dB
mean ΔMCD	+3.35 dB
median ΔSTFT error	+0.03 dB
sign test, win count	$p = 3.2 \times 10^{-6}$
Wilcoxon, ΔMCD	$p = 6.4 \times 10^{-12}$
Wilcoxon, ΔSTFT	$p = 1.4 \times 10^{-9}$

Every sign is the opposite of the respiration arm's, and the same pre-registered prediction — win share below 50 % with a median ΔMCD at or below zero, made before the run and not adjusted after watching respiration lose — **failed**. Both metrics improve, and on a population large enough that a 5.6-point majority is not a coin: the sign test and the Wilcoxon both reject at the margins above, and the two metrics happen to return the same win count from different type sets (294 types win on MCD alone and 294 on STFT alone).

The size of the effect should not be oversold. A me-



Left: pale and thin, Pearson; solid and thick, Spearman. Neither of the cosine's correlations is distinguishable from zero and both point estimates lean the wrong way. Not an instrument artefact: the recomputed cosine reproduces the stored one at Pearson 0.921 and the two distortion metrics agree with each other at 0.925. Right: blended into the objective at weight 0.5 the mel target wins 25 of 80 types, with a median ΔMCD of exactly 0.0 dB and a maximum cosine loss of 0.120; seven of ten had won at probe scale. $n = 200$ types then $n = 80$; both panels are debug-scale.

Figure 10: Left: correlation with audible distortion for the incumbent objective and its proposed replacement, $n = 200$ types on the 100-species debug store against the cached audio. Right: the replacement's win share at exploratory scale and at pre-registered gate scale. Both panels are debug-scale, and both are *superseded*: the table above this figure carries the same comparison at $n = 1\,200$ on the production store, where the incumbent's correlation with distortion becomes significant and keeps the wrong sign.

dian of +1.13 dB MCD and +0.03 dB STFT is small, the distribution is wide in both directions (10th percentile −11.2 dB, 90th +20.1 dB), and the mean is carried by a helpful tail. The claim is that this adoption is significantly *not* harmful, not that it is worth much.

So the caveat is real and it is not a law. One measured instance of a cosine-gated adoption went the wrong way and one went the right way, and the population sizes run against the comfortable reading: the arm that held is the larger by a factor of ten. A proxy that is anti-correlated with what you care about *across* types does not have to damage every decision it gates; it removes the guarantee that it does not. Which of the two outcomes a given adoption got could not be predicted from the proxy, and was not.

The remaining two cannot be paid without a refit, and the reason is on disk. The fricative route is not reversible by stripping its field: `fricative_seed` re-searches the noise axis and re-seeds the formants on the winning path, so a shipped fricative type no longer carries the values it had before the route was taken, and removing the flag alone renders fricative-tuned parameters through the tonal path rather than reconstructing the control. The pre-adoption snapshot that would settle it does exist (`data/learned-inventory-pre-fricative.json`), and it cannot be used, because the typing it was learned from was overwritten by the current production run: benched against the current store it skips 82 of 112 types for having no members and recomputes cosines that disagree with its own stored ones, which is the wrong-store signature. `docrefit` ships on 10 types in the current inventory, which is below any population worth a verdict. Both remain owed, and both now have a named cost — a refit with the stage disabled — rather than an open question.

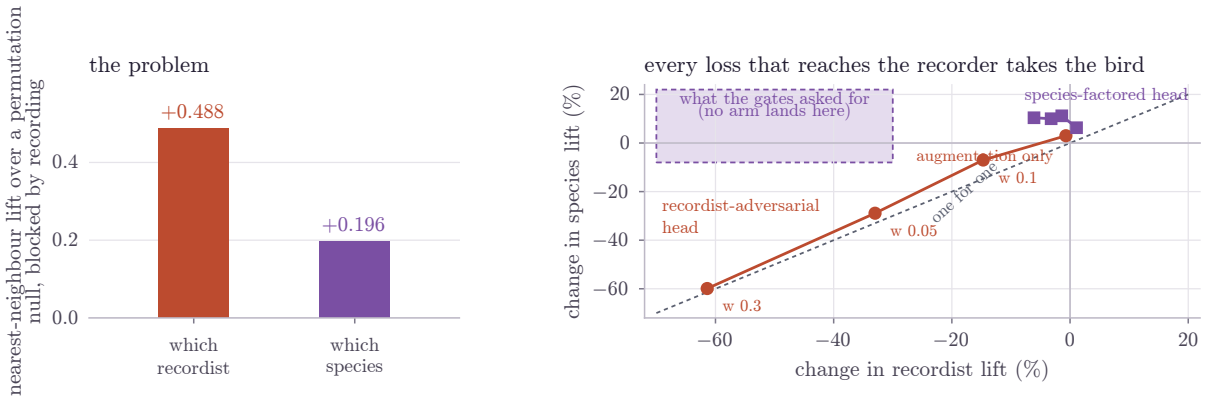
7.3 Limit 2. The embedding sorts by recordist, not by bird

In the critic's embedding, the strongest relation is not the animal. Nearest-neighbour lift over a permutation null, blocked by recording, is +0.488 for recordist and +0.196 for species. Every consumer of the embedding inherits that: typing, the map, the relations document, novelty and the unknown-bird path.

The road out has now been closed at four independent ends, which is Section 8's and Section 9's subject and is summarised here. Training-time augmentation that simulates a recorder's frequency response moves the recordist lift by −0.7 %, i.e. not at all. A gradient-reversed recordist head [23] does move it, at an almost exactly one-for-one cost in species structure. A species-factored version of that head protects species completely and reaches only −6 % of the confound. Post-hoc batch correction, including the strongest member of the genomics toolkit [38], reproduces the same one-for-one exchange rate by a completely unrelated mechanism. And two archive-scale external models, neither of which has ever been asked to tell our resynthesis from a bird, still carry between five sixths and more than all of the recordist structure that ours does.

The reason is now visible and is a statement about the sample, not about any model: most species in this corpus are carried by a handful of recordists, so no method that sees only the embedding can separate labels the data never separated. Subtracting each recordist's mean leaves the recordist lift *higher* than it started (0.1665 to 0.1687) while costing a third of the species structure, so the confound is not a location shift and not a per-dimension gain either. The named action is more recordists per species, in the data.

That action now has a threshold, which is the one place in this paper where a limit moved. Measuring the cost of a recordist-blocked split separately by how many recordists carry the species — label-set size equalised at 115 species a stratum, recordings



Left: in the critic's embedding the strongest relation is not the animal, and every consumer inherits it - typing, the map, the relations document, novelty, and the unknown-bird path. Production store. Right: every swept point of both adversarial heads. The shaded region is what the pre-registered gates asked for and no arm reaches it; the diagonal is a one-for-one exchange of recordist structure for species structure, and the plain head tracks it. The species-factored head protects species completely and reaches only 6 % of the confound. 37-species debug store; debug-scale.

Figure 11: Left: what the embedding organises by. Right: every swept point from both adversarial heads, as relative change in recordist lift against relative change in species lift. The shaded region is what the pre-registered gates asked for and no arm reaches it. 37-species debug store; debug-scale.

per species flat between 8.2 and 10.9 — gives 30.82 points of drop at two recordists, 20.92 at three, 18.52 at four, 11.12 at five and **3.08 at six or more**. A factor of ten, monotonic. So the slogan resolves to a number: at six independent recordists a species is close to safe under a random split, and at two it is not usable for benchmarking at all without blocking. It is observational rather than causal — species are not assigned recordists at random — but it is the only intervention we have found that removes this confound instead of measuring it, and it belongs to the archives rather than to any model.

And the confound is those recordists, individually, in both external models. Per recordist findability — how much more often a recording's nearest neighbour is also theirs than chance allows — correlates between Perch 2.0 and BirdNET V2.4 at a partial r of $+\mathbf{0.852}$ over 212 recordists, after the number of recordings each contributed is removed from both sides, against a shuffled null 95th percentile of $+0.140$. 83 % of recordists are findable in both spaces. Two architectures trained by different groups agree about *which people* are identifiable, which locates the structure in the recordists' equipment, sites and habits and not in either model's inductive bias.

One thing this limit does *not* reach, and it is worth as much as the rest. The question a comparative ornithologist would ask is whether the signature reaches the *estimate* — the single number per species that every cross-species acoustic comparison built on an opportunistic archive rests on. It does, and it is large. Over 797 species with two or more recordists contributing at least 20 syllables each, the within-species between-recordist variance in median $\log f_0$ is 0.0446, against 0.0033 for a null that shuffles recordist labels *inside* each species — twelve times the null, and **0.638 of the variance that separates different species**. Two people recording the same bird disagree about its pitch by nearly two thirds of what separates one bird from another.

And the comparative slope survives it. Estimating the mass/frequency relationship the standard way (one median over all of a species' syllables) gives a slope of -0.1655; estimating it recordist-balanced (each recordist one vote, then the median across them) gives -0.1672, a change of **1.0 %** against a pre-registered tolerance of 15 %. A recordist effect that is large but close to mean-zero across species *attenuates* a slope rather than tilting it, and the supporting correlations agree: a species' recordist count correlates with its naive f_0 at only $+0.125$. This is the reassuring half of the limit, and it also names a specific component of the noise behind two numbers this paper already reports — G2's standardised slope of -0.211 against the published -0.510, and a fitted Pagel's λ of 0.000 that cannot be true of birdsong. The population is also an order of magnitude larger than the one G3 could reach, because this test is a direct measurement rather than a fitted parameter and it is paired *within* species, so the selection that binds G3 barely applies.

Two earlier measurements sharpen the same limit. Both are reported in full in Section 12.1. **The structure is not an artefact of any model having memorised these archives**: on 2 158 recordings uploaded to xeno-canto after Perch 2.0's training snapshot, that model's recordist lift is $+0.441$ against a species lift of $+0.750$, both scored on identical rows at full coverage. And recordings per species correlates with distinct recordists per species at $r = +0.923$ over 3 853 species, with a median of 17 recordists for species holding at least eight recordings against 2 for the rest. That correlation is also a warning about test design, because any analysis requiring several recordings per species selects the part of the corpus where this limit binds least.

7.4 Limit 3. Syntax is bounded by the data and not by the method

A first-order transition model over the learned tokens is strongly supported (Section 6). A second order is not, and the reason is countable: of 2 421 species, 217 clear the 20-bout bar at which the question can be asked at all. Within those, the median species *loses* 0.091 bits per token going from order 1 to order 2 on bouts it has not seen, at three independent splits. The literature is unambiguous that birdsong is not first-order [34, 44], so this is a statement about the corpus, not about birds. The depth harvest is the action, and it has a hard ceiling: the tenth percentile of the 100 deepest species is 5 recordings, which is what the archive holds.

7.5 Limit 4. The map confuses its own regions

Two ceilings sit on the claim "sing, and the right region lights".

The first is in the embedding. With every type granted its own cluster, so that no type can fail to be retrieved for lack of a cluster. The share of labelled points whose nearest centroid by cosine belongs to their own type is 0.602. For 40 % of points the nearest centroid is another type's. That is the embedding's confusion between types, not the clustering's, and it was invisible until label assignment removed the competing explanation.

The second is in the bridge. On held-out recordings, the share of syllables whose nearest region is the same through the ridge map as through the critic itself is 0.61 on the debug build and 0.66 on production. Decomposing 1 439 flips over 3 731 held-out syllables locates it: 76.4 % is boundary near-ties, with 74 % of all rows sitting inside a margin of 0.05 and 41 % of flips landing on the runner-up region. Genuine large-margin map failure is 4.3 % of the loss. Tie-adjusted agreement is 0.91 and species-level agreement is 0.781 against chance 0.033. The consequence adopted was to make the single-label claim at species level and to label region-level claims as "nearest of several close regions"; no better map class was pursued, because the addressable loss is at most 2.6 % of rows.

7.6 Limit 5. The failure is in the fit, not the oscillator

An audit against the canonical literature found no deviation inside the oscillator equations. The implemented ODE is the published normal form term for term, validated against Runge–Kutta, against the reference oscillator measurements, against a two-sided parity harness, and against the literature's own stated coupling-approximation boundary. The two-voice kernel's measured collapse into period doubling occurs exactly where Laje and Mindlin [41] says the weak-coupling approximation dies.

Every recorded deviation sits outside the oscillator, and the list is specific. A single f_0-tracking band-pass makes the spectral envelope pitch-invariant: measured relative partial gains span 1.3-fold over a 5.7-fold change in f_0, against 32-fold for a fixed waveguide, so the structure locks spectral shape to pitch. Pinning every syllable to $\beta \approx 1$ places it at the tonal end of the model: the resynthesis reproduces the median spectral content index almost exactly (1.099 against 1.092) and has essentially no rich tail (1.4 % above 1.5 against 15.6 % of real syllables; 0 % above 3 against 5.3 %), with median spectral flatness three orders of magnitude too pure at 0.0005 against 0.044. The tension clamp bites in 6 617 of 12 724 types, so over half the corpus has contour excursions riding a parameterisation limit while audible. Pressure arrives as a designed envelope and not as respiratory physics. And there is no physiological micro-variation: real crystallised song carries about 1 % rendition-to-rendition fundamental-frequency variation [72], and the implemented roughness mechanism is one static fitted depth per type.

Two reading notes follow. First, "noisy" and "machine-like" are opposite failure modes of one cause family: the representation has one noise axis and one envelope, so a fricative collapses to unshaped hiss and a tonal note to a metronomic sine. Real syringes live between the regimes, and the model contains that middle. Period-doubled and chaotic solutions are in the equations [1, 68], but the fit cannot find its way there under the current objective. Second, the diagnosis is now sharper than "the fit fails", because the starting point has been ruled out: an amortised inverse encoder that proposes a full parameter vector in one forward pass reaches a comparable basin about as often as the cluster median already does (Section 8). What remains is the objective and the physics beyond the oscillator.

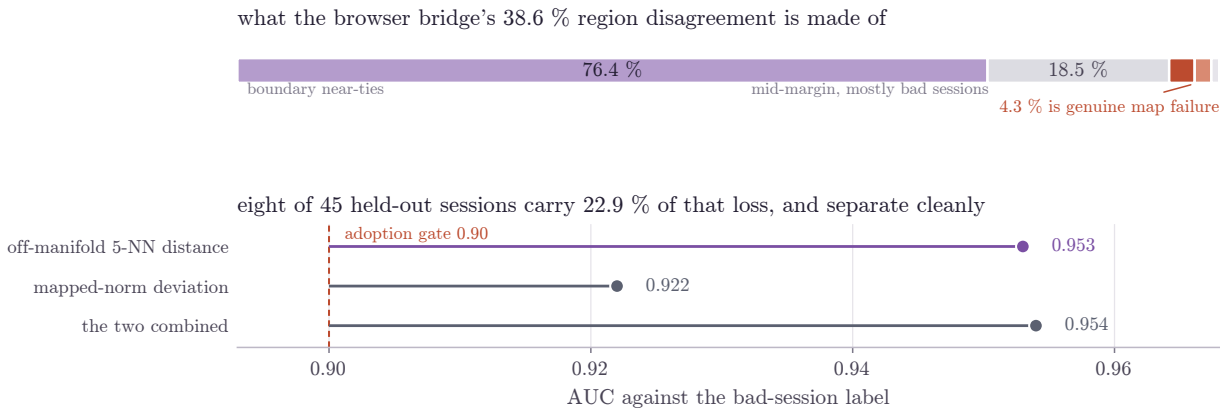
8 Our methods against each other

This section reports the head-to-head arms the project ran against itself. Each states what it licenses and what it does not. Unless noted, arms sharing a table were run on one fixed corpus or one fixed dumped matrix, through the shipped functions, not through re-implementations, so they are comparable to each other and not necessarily to arms in other tables. Throughout, "points" means percentage points: a retrieval score of 0.5775 rising to 0.6020 is +2.5 points.

8.1 Hand-weighted typing against the learned embedding

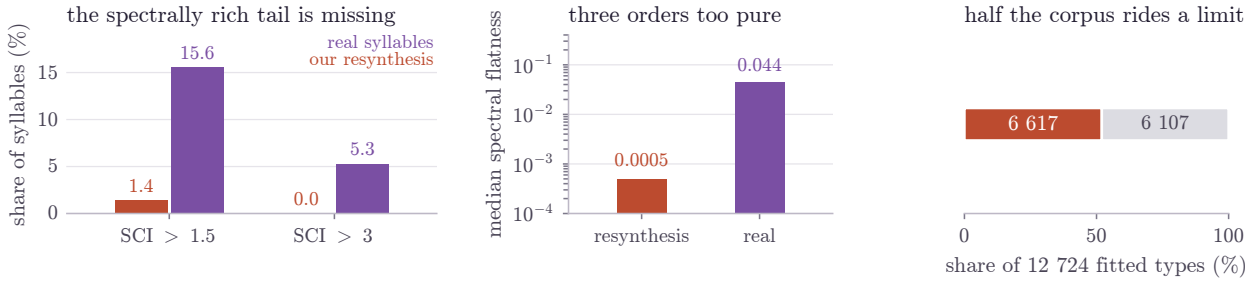
Syllable identity was originally decided by seven hand-weighted numbers. Replacing that vector with the critic's 32-dimensional embedding is the largest accuracy gain the project has measured. The headline metric is species placement: does a resynthesis land nearest its own bird's centroid, not another species'?

Table 1 licenses three statements and refuses a fourth.



Top: 1 439 flips over 3 731 held-out syllables. Tie-adjusted agreement is 0.91 and species-level agreement, which is the honest single-label claim, is 0.781 against chance 0.033; the addressable loss is at most 2.6 % of rows, so no better map class was pursued. Bottom: the browser can compute the winning signal from the 28-d side alone, and a shipped cloud of 256 random anchors holds AUC above 0.92 for 28 kB. Using it as a lever - raising the remaining rows' agreement by 5 points - failed at 3.5, so the pre-registered fallback fired and it ships as an honesty gate. Debug build.

Figure 12: Top: the 38.6 % region disagreement between the browser's bridge and the critic, decomposed over 3 731 held-out syllables. Bottom: separation of the eight pathological sessions by each candidate live signal, against the pre-registered adoption gate. End-to-end debug build; debug-scale.



Pinning every syllable to $\beta \approx 1$ places it at the tonal end of the model. The median spectral content index is right - 1.099 against 1.092 real - and the tail is not there at all: the model contains the pulse-like regime that produces it and the fit does not reach it. Median spectral flatness is three orders of magnitude too pure, which is a whistle too clean to be an animal. Right: the tension clamp bites in 6 617 of 12 724 types while the syllable is audible, so over half the corpus has contour excursions riding a parameterisation limit rather than a fitted value.

Figure 13: The tonal-regime pin, in three panels. Left: share of syllables above two spectral-content thresholds, resynthesis against real. Centre: median spectral flatness on a log axis, three orders too pure. Right: the share of fitted types whose tension contour rides the clamp of the parameterisation — 6 617 of 12 724, so half the corpus sits at a limit of the parameterisation rather than at a fitted value, which is a different failure from the first two and is the one that bounds them. The model contains the pulse-like regime that produces the missing tail; the fit does not reach it.

Learned typing is a clean win, replicated on two stores at two species counts: +6.3 points at 9 k and +13.3 points at 40 k, with the guardrail cosine essentially held and pitch error improving in both. Nothing regresses.

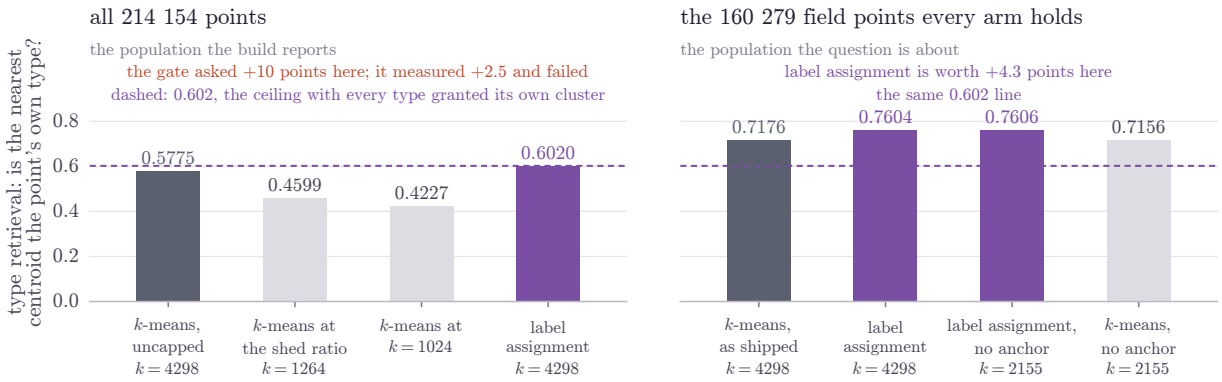
The learned realism logit as a *fit objective* is refuted, and it fails in exactly the predicted way: monotonically harmful in its weight, with the cosine collapsing from 0.762 to 0.335 at $w = 1$ while placement drops to 17.3 %. A critic at realism AUC 0.990 on entirely held-out species rules out "the critic was not good enough". Separating two distributions is not the same as having a gradient that points at the nearer one. The guardrail term earned its place by catching the thing it was put there to catch, and the default weight is zero.

The own-type placement column, which we omit from the table, moves the wrong way as species placement rises, and that is an artefact, not a regression: the learned embedding drove the type count from 138 to 294, and own-type placement asks whether a resynthesis beats *every other type of the same bird*. With twice as many rivals it is a strictly harder question, and it is not comparable across rows with different type counts. That is why species placement is the headline.

What the table does *not* license is the waveguide row. A structurally fixed tract with 32 times the spectral freedom of the tracking band improves all three metrics at once on five species and was a regression across 22 species on a different store; here it wins under both typings and pays for it in cosine, missing the pre-registered guardrail. It remains promising and unconfirmed, and it is gated off.

8.2 Two spaces: the 28-dimensional descriptor against the 32-dimensional embedding

The two spaces are not interchangeable and the choice is measured, not assumed. On a controlled subset, 625 field syllables, three species, identical points, one seed,



The question is the one the browser asks of live audio: is the nearest centroid by cosine the point's own type? One fixed matrix dumped from a 100-species debug build - 214 154 points, 2 143 types, 160 279 of them from field recordings - with every arm scored through the shipped clustering, seeding and purity functions. On a corpus that fits under the cluster cap, seeded k-means is nearly as good as knowing the answer, which is why the gate fails and why the failure is the useful part. Changing the query population moves the same arm by 14 to 16 points, more than any arm moves it, and an arm that drops the synthetic copies drops the hard queries, so on all points it flatters itself by 18. Purity and coverage are identities under label assignment, not measurements. Debug-scale.

Figure 14: Type retrieval against the cluster budget, and against the query population. The question is the one the browser asks of live audio: is the nearest centroid by cosine the point's own type? Left: scored over all points, label assignment beats the shipped k-means by 2.5 points against a pre-registered 10, and fails. On a corpus that fits under the cluster cap, seeded k-means is nearly as good as knowing the answer. Right: scored on the field points every arm holds unchanged, which is the population the question is about. The query population changes the same arm's answer by more than any arm changes it. Debug-scale.

Table 1: Typing space and fit objective, ablated on two stores. Each row changes one thing. *Species* is the share of resyntheses landing nearest their own species; *cosine* is the guardrail the critic cannot touch. Absolute values are not comparable between the two panels: the stores differ, and the right-hand panel is a 13-species measurement after a store-trim bug deleted twelve species, so chance there is 7.7 % and not 4.0 %.

	9 041 syllables, 25 species				39 993 syllables, 13 species			
arm	types	species	cosine	$\lvert f_0 \rvert$	types	species	cosine	$\lvert f_0 \rvert$
hand typing, cosine objective	346	26.6 %	0.762	5.4 %	138	38.4 %	0.807	3.3 %
+ realism objective, $w = 0.25$	346	27.2 %	0.686	6.6 %	(	)	(	)
+ realism objective, $w = 0.50$	346	23.7 %	0.655	6.5 %	138	31.9 %	0.642	4.0 %
+ realism objective, $w = 1.00$	346	17.3 %	0.335	5.7 %	(	)	(	)
learned-embedding typing	319	**32.9 %**	0.755	3.1 %	294	**51.7 %**	0.796	2.2 %
hand typing + waveguide tract	346	30.1 %	0.719	5.4 %	(	)	(	)
learned typing + waveguide tract	319	37.0 %	0.706	3.6 %	294	57.1 %	0.723	2.1 %

only the space changed. Moving the build into the embedding takes cluster purity from 70.9 % to 90.9 %, type coverage from 12/14 to 14/14, three-dimensional 10-nearest-neighbour agreement from 64.5 % to 83.2 % and distance faithfulness from 0.874 to 0.940. On the 100-species debug corpus the descriptor-space discovery map degenerated to three clusters with one holding 95 % of points. INaturalist soundscape audio is one density blob in 28 dimensions. Where the embedding gives 98 clusters with real structure.

Two counterweights sit against that. A render-free placement check scores the 32-dimensional space slightly *below* the 28-dimensional one (38.5 % against 41.3 %), and the recordist confound of Section 7 lives in the embedding, not in the descriptor. The embedding is the better similarity space and the worse-behaved one.

8.3 Objectives: cosine against a mel-path target

Reported above in Section 7 and summarised here for the comparison table. The mel target is the better *measurement* and the worse *optimiser*: it ranks audible distortion where the cosine cannot (Spearman −0.420 against +0.100), and as an objective at $w = 0.5$ it wins 31 % of 80 types against a 60 % bar with a median effect of exactly zero. The two objectives agree only moderately on which types are good (Pearson 0.456), so they are complementary and not redundant. An objective that is both a good ruler and a good optimiser remains unfound.

8.4 Clustering: seeded k-means against label assignment

Ninety-six per cent of a production corpus's points already carry a *(species, syllable)* label by construction. The default build discards it, partitions the space geometrically, and asks each partition to vote a name back

Left: 625 field syllables, three species, identical points, one seed. Hollow marker, the 28-d descriptor; filled, the 32-d embedding. Clay is the counterweight, because render-free placement moves the wrong way and the recordist confound lives in the embedding rather than in the descriptor: the embedding is the better similarity space and the worse-behaved one. Right: the browser computes only the 28-d descriptor, so a kernel ridge over random Fourier features ($D = 2048$, $\sigma = 4$, $\lambda = 1$) maps it into the 32-d space the map lives in. It is gated on held-out recordings, never on held-out syllables, because two syllables of one bird in one take are nearly the same pair twice; it ships as plain numbers, and no weights reach the browser. Both panels are debug-scale.

Figure 15: The two feature spaces and the bridge between them. The browser computes only the 28-dimensional descriptor; the map lives in the critic's 32-dimensional embedding; a kernel ridge over random Fourier features maps one into the other behind a ship gate on held-out recordings.

by point-count majority. Label assignment takes the other side of that design decision: every labelled point goes to its own type's cluster and k-means fits only the unlabelled remainder.

All arms in Table 2 run on one fixed matrix, 214 154 points, 2 143 types, 160 279 of the points from field recordings. Dumped from a 100-species debug build, and every arm is scored through the shipped clustering, seeding and purity functions. The harness was validated twice before any arm was read: the $k = 1024$ row reproduces the earlier baseline exactly, and the $k = 4298$ row reproduces the shipped debug build's own manifest.

The pre-registered gate asked for +10 points of type retrieval against the shipped k-means build. It measured +2.5 and **failed**. The failure is the useful part. The mechanism the gate was predicted from is types shedding their only cluster, and under the raised cluster cap this corpus wants 4 298 clusters and gets all 4 298, so that mechanism never fires. *On a corpus that fits under the cap, seeded k-means is nearly as good as knowing the answer.* A real result about the seeding, and the argument against making label assignment the default. Placing this corpus in production's shedding ratio ($k = 1264$) and at $k = 1024$ gives +14.2 and +17.9 points. The scope of the claim is therefore exact: label assignment is worth what the cluster cap costs, and nothing more.

Two of the build's headline numbers stop being measurements under this arm. Purity 0.9988 and type coverage 1.0000 hold because a type *is* a cluster here, by construction. The shipped manifest says so in a dedicated field, because 1.0 read as a good map is reading the definition back. Type retrieval is the statistic that survives, because nothing forces it.

There is a measured cost, and it is in the live path. Against the k-means build of the same corpus, region-level agreement through the bridge falls from 0.6572 to 0.6316 and species-level agreement from 0.7421 to 0.6831, while the tie-adjusted figure *rises* from 0.8032 to 0.8136. Read together that is more near-ties, not worse decisions: centroids placed at label means are less separated than centroids Lloyd's iteration pushed apart, so more live sounds land between two regions, and the ones that land decisively land better. A build whose headline is the species claim should read that row first.

8.5 The discovery mode at production scale

The clustering arms above are two ways of naming everything. The build's third mode is the only one that can decline: HDBSCAN* finds however many dense regions the corpus holds and leaves the rest unassigned, so a sound can come out as *none of these*. Every number reported for it until now was a debug-scale extrapolation. This is the production measurement, and it is a negative result with one genuinely encouraging half.

The matrix is the one the build clusters, rebuilt from the harvest store's 32-dimensional embeddings and the build's own standardisation moments: 435 630 of the shipped 444 213 rows (98.07 %). The missing 8 583 are rendered anchors whose embeddings exist only inside a build, so the reconstruction is checked against the build's own manifest before anything else is read from it, and it holds — 1 128 regions and 79.3 % unassigned here against the build's 1 151 and 79.5 % on a corpus 1.93 % larger.

The refusal is not a parameter that can be tuned away. Raising the minimum cluster size from 25 to 100 does cut the unassigned share from 79.3 % to 36.8 %, and that is the only thing about it that improves. The number of regions falls to **two**, the larger holding 275 175 of 435 630 rows, and agreement with the label partition over the rows it places falls to nothing (ARI 0.0000, AMI 0.0006). It has stopped distinguishing anything. So the honest report of the discovery mode on this corpus is that it declines four syllables in five at the only setting that still separates them, and the alternative is not a coarser map but a

Table 2: Clustering arms on one fixed 100-species debug matrix (214 154 points, 2 143 types). *Type retrieval* asks, for each point, whether the nearest centroid by cosine, the lookup the browser performs on live audio, belongs to that point's own type. Purity and coverage are identities under label assignment, not measurements. Debug-scale.

arm	k	purity	type coverage	type retrieval	clustering (s)
k-means, uncapped (as shipped)	4 298	0.5797	0.9258	0.5775	64.9
k-means at production's shed ratio	1 264	0.4601	0.5114	0.4599	25.5
k-means at $k = 1024$	1 024	0.4232	0.4260	0.4227	27.2
label assignment	4 298	*0.9988*	*1.0000*	**0.6020**	**0.2**

Table 3: The discovery mode on the production corpus (435 630 rows, 32-d critic embedding). Agreement is against the shipped 15 341-knot label partition and is computed *twice*: over the rows this clustering placed, and over every row with *unassigned* as a label of its own. At four fifths unassigned the two ask different questions, and reporting either alone misleads. Production scale.

			cluster size		placed rows only		all rows
min. cluster	regions	unassigned	median	largest	ARI	AMI	AMI
25	**1 128**	**79.3 %**	58	719	**0.4436**	**0.8013**	0.1580
100	2	36.8 %	137 652	275 175	0.0000	0.0006	0.0434

single lump. A reader tempted to read 79 % as a tuning failure should read the second row.

Where it does commit, it and the labels are describing the same thing. This is the half that changes the interpretation, and it is why the table reports agreement twice. Over all rows, adjusted Rand is 0.0001 and adjusted mutual information 0.1580 — the figures of two unrelated partitions, and the ones a single class holding 79.3 % of the points will produce whatever the structure is. Over the 90 193 rows the clustering actually placed, adjusted mutual information is **0.8013**. The dense regions are therefore not a coarsening of the label knots, because they do not cover the corpus; and they are not independent of them either. The refusal is a coverage property, not a disagreement about what the corpus contains.

And the refusal is not an artefact of recording quality, which is the first alternative explanation to reach for and the one that would make the whole discovery surface dismissible. Two independent channels say otherwise. Neighbour agreement — the share of a point's ten drawn neighbours sharing its type, computed from a different array and on the *other* map's positions — is 0.8505 among placed points and 0.3562 among refused ones, separating the two classes by a factor of 2.4. Signal-to-noise ratio barely moves: a median of 18.67 dB for placed points against 17.27 dB for refused, 1.4 dB apart on a distribution spanning 2.14 to 39.34 dB between its fifth and ninety-fifth percentiles. So "the density estimate declined it" does not mean "the audio was bad", and an SNR floor would not clean this map up.

The consequence adopted is the narrow one. The discovery map is not shipped as the map: its projection faithfulness (Pearson 0.3016 against 0.429) and neighbour agreement (0.2945 against 0.4574) are both materially worse than the label build's, and 14 323 of 15 241 types have no region carrying their name, so they could not be lit, labelled or given a trajectory in it. What ships from it is one column — whether each point fell in any dense region at all — which is what lets the interface mark the four fifths of the corpus that no density estimate will group, rather than implying that every drawn point sits in a region something has claimed.

One reproducibility note that cost a working command to find. HDBSCAN* here is one algorithm with two minimum-spanning-tree implementations: scikit-learn's Prim's algorithm, and the standalone package's Boruvka tree, which is the only one that finishes at this scale (roughly 22 minutes against an extrapolated 8.4 hours). They are widely treated as interchangeable and at this scale they are. *At pool scale they are not*: on sixteen observations at a minimum cluster size of eight, scikit-learn returns two clusters and no noise where the standalone package returns one cluster and eight noise. Preferring the fast implementation unconditionally therefore changed what the unlabelled-arrival path discovers, silently, and the symptom appeared as a self-test failure rather than as a clustering difference. The implementation is now selected by row count. Anyone swapping an MST implementation on the strength of "same algorithm, same labels up to ties" should measure the small-n case.

8.6 The synthetic member anchor ("W5")

The map was built out of both real and synthetic points: each type contributed one rendered anchor plus renders of up to 24 measured member realisations, alongside the field syllables those members already are. At production scale that made 51.1 % of the map a rendered copy of the other half (347 531 synthetic against 340 079 field). Because the critic's leading loss term is realism, the recorded/synthesised axis is the strongest relation in its embedding (lift 0.990 against species' 0.513), so every type lived in two blobs and the cluster budget bound at twice the type count.

Three arms were run on the fixed matrix. Reducing member variants to the anchor alone moved seeds per type from 2.000 to 1.993, i.e. not at all. The source split

is *binary*, splitting whenever a type holds points on both sides, so 24 renders and one render split it equally. Deflating the source axis worked exactly as specified, removing 42.0 % of the space's energy and driving the gap between the two source means from 7.233 to 0.0, and *every type still split* (1.999), with purity falling 1.1 points. So the recorded/synthesised difference is not one direction: the global offset is, and the per-type geometry the seeding gates on is not. What works is rendering nothing at all for a type that has member observations, since it already exists in the map as the real syllables the typing assigned to it: seeds per type falls to exactly 1.000 and the wanted cluster count halves from 4 298 to 2 155.

The follow-up changed what that arm is *for*, and it is the cleanest methodological lesson in this section. Type retrieval over *all* points is not comparable between arms that hold different points: dropping the synthetic copies drops the hard queries, so the arm flatters itself by 18 points. Scored instead on the 160 279 field points every arm holds unchanged, which is the population the question is about, since a recorded syllable must light its own region. The picture is different.

The anchor was removed on the size argument and not the accuracy argument, and the decision is recorded as such. It halves the region count and removes a quarter of the points, taking the production build from a wanted 27 862 clusters to 14 003 and the centroid asset from a projected 44 MB to 7.69 MB of JSON plus 3.81 MB of binary sidecar. The difference between "needs a new asset format" and "ships". It costs the rendered voice its own place on the map, which is a claim about our synthesiser, not about a bird. One number moved the wrong way and it is a picture number: three-dimensional distance faithfulness fell from 0.903 to 0.747 while 10-nearest-neighbour agreement rose from 20.7 % to 28.0 %. Removing 53 575 rendered points removed structure the projection found easy, so distance on screen means less than it did. Neither the regions nor the matching ever used those positions.

8.7 Where the descent starts: cluster median against an amortised inverse

Coordinate descent starts each type's gesture axes at their defaults, which are wrong for 78 % of types on `gammaScale` alone. Since we own the forward model, training data for an inverse is free: sample a parameter vector, render it, and measure the render with the same patch function the critic consumes. This is the encoder-decoder arrangement of differentiable DSP [17] with a physical decoder, which is why the interpretability survives.

The gate was written before the encoder was trained, and both predicted weak spots were recorded before the run: `tractQ` is the worst-learned head by an order of magnitude (validation MSE 1.125 against `gammaScale`'s 0.102) and timbre classification reaches 0.442 over three classes where chance is 0.333.

The median and the mean both rise, so the encoder does something real; it improves barely more types than a coin would, and it costs one species a tenth of its fit. It is not adopted, and the flag is off by default.

The finding that matters is diagnostic. *The starting point was not the binding constraint.* Coordinate descent from the cluster median already reaches a comparable basin about half the time, which is more than the plan assumed when it described the search as unable to escape a bad basin. So the remaining fit error is mostly not a basin problem, and what is left of "the model is right and the fit is what fails" lives in the objective and in the physics beyond the oscillator. The mechanism of the encoder's losses is visible in a separate arm: an encoder-only start scores *below* the cluster median it replaces (0.342 against 0.442), because on the fields the median already measures. Timbre, attack, hold, harmonic, the modulations. A regression estimate is worse than a measurement. Its only new information is on the four gesture axes the median carries no value for, and the wins cluster exactly there.

A two-start variant was pre-registered in the same file, not re-running the gate with a softer bar: run the descent from both starts and keep whichever scores better on the objective that already decides, which cannot regress by construction. Its bar is +0.02 median with *no* species worse at all. Its measured cost is roughly 61 minutes against the single-start arm's 11, because the encoder start explores a different basin and its descent does not terminate as early.

It passes, and it is the only pre-registered gate the inverse encoder has cleared. Median resynthesis cosine rises 0.7410 → 0.7740, a gain of +0.0330 against the +0.02 bar; 51.7 % of the 1 661 shared types improve, **none regress**, and no species median falls (87 up, 13 unmoved, 0 down). Debug-100 subset, scored through the fit's own per-type log against the baseline arm.

The route there is worth reporting, because the gate's second condition earned its keep. The first two-start run cleared the median bar at +0.0320 and *failed* on 48 of 1 661 types coming out worse, by up to 0.049. Since a two-start cannot regress by construction, the pre-registration had already named that outcome a bug in the selection and not a trade, so it was diagnosed, not excused. Three candidate explanations were eliminated by measurement. The search objective and the reported metric are the same computation; the fricative refinement is monotone in its own input; the two arms fitted identical inventories. A fourth check was direct: the renderer is bit-identical for fixed parameters, so the reported cosine is a value and not a draw. The cause was that the selection compared candidates *before* the fricative refinement while the shipped and gated number comes *after* it, and that refinement's benefit depends on the basin: 19.3 % of types carry the fricative flag, and 100 % of the 48 regressions did. Scoring each candidate as it would ship took the regressions to zero and the median gain from +0.0320 to +0.0330.

Two limits belong with the number. It licenses a fit option that is never worse; it does not license an audible claim, because the gate reads the same 28-d

Table 4: The same four arms scored on one query population: the 160 279 field points every arm holds unchanged. Label assignment is worth +4.3 points and is the same in both anchor conditions, so the effect is the clustering's. Removing the synthetic anchor is worth nothing on this statistic; what it is worth is half the regions and a quarter fewer points. Debug-scale.

arm	points	k	field-point type retrieval	clustering (s)
k-means, as shipped	214 154	4 298	0.7176	54.8
label assignment	214 154	4 298	**0.7604**	0.2
label assignment + no synthetic anchor	160 579	2 155	**0.7606**	0.1
k-means + no synthetic anchor	160 579	2 155	0.7156	15.2

Table 5: The amortised inverse encoder as an initialiser, on two `learn` runs over the 100-species debug store differing only in one flag. 1 661 fitted types, all shared. The gate was stated in full before the encoder was trained and fails on all three conditions. Debug-scale.

	baseline	encoder	gate	
median resynthesis cosine	0.7411	0.7546 (+0.0135)	$\geq +0.02$	**fail**
mean resynthesis cosine	0.6824	0.6997 (+0.0173)	,	
share of types improved	,	877/1 661 = 52.8 %	$\geq 60\,\%$	**fail**
worst species' median change	,	−0.0816	≥ -0.05	**fail**

cosine whose correlation with spectral distance is about 0.1 (§8.3), so a bench run on cached audio is what an audible improvement would need. And the single-start arm's published worst-species figure of −0.0816 was scored through the same pre-refinement comparison, so part of it is that artefact, not the encoder's fault, which does not rescue that arm, since it failed on the median alone.

8.8 Grammar order on held-out bouts

Four models, one protocol: per species, an 80/20 split by *bout*, a k-order backoff model fitted on the train half for $k = 0 \ldots 3$, every held-out token scored, cross-entropy in bits per token. The corpus is 34 696 bouts, 327 164 tokens and 2 421 species, of which 217 clear the 20-bout bar. The control is the same corpus with each species' tokens redealt within species, keeping bout lengths and marginals. The loader is pinned against the shipped bout function and refuses to run if the bout rule has drifted.

This table is the source of both the strongest positive result in the paper and one of its clearest negatives, and Section 6 reads it in full.

A separate bake-off compared model *classes* at first order and above. A repeat-count-dependent model in the style of Jin and Kozhevnikov [34] lost decisively at −0.348 nats per token, winning 0 of 30 species. A hidden-state HMM won the pre-registered rule at +0.422 nats and 29 of 30 species, and the decomposition reversed the adoption. Splitting held-out tokens by whether the transition was seen in training, first-order is the better model on the 90 % seen (−1.525 against −1.624 nats/token, better in 15 of 30 species), and the HMM's entire advantage is the 10 % unseen tail, where first-order pays −9.25 and the HMM −4.16. That is a property of smoothing *shape*, because a mixture predictive is heavy-tailed. It is not evidence of recovered hidden structure. Adopting a heavier model for a smoothing artefact would be the wrong lesson, so first order stayed and the smoothing was repaired instead: a Kneser–Ney continuation backoff [37] gains +1.90 and +2.32 nats per token on unseen-*context* transitions across two debug stores with the seen side bit-identical, and passes its own separately pre-registered gate. Two caveats: 80 % of the unseen tail is seen-context with an unseen continuation, outside the repair's scope, so the full-tail gain is +0.37 and +0.65; and the repair narrows the HMM's advantage, which confirms the diagnosis.

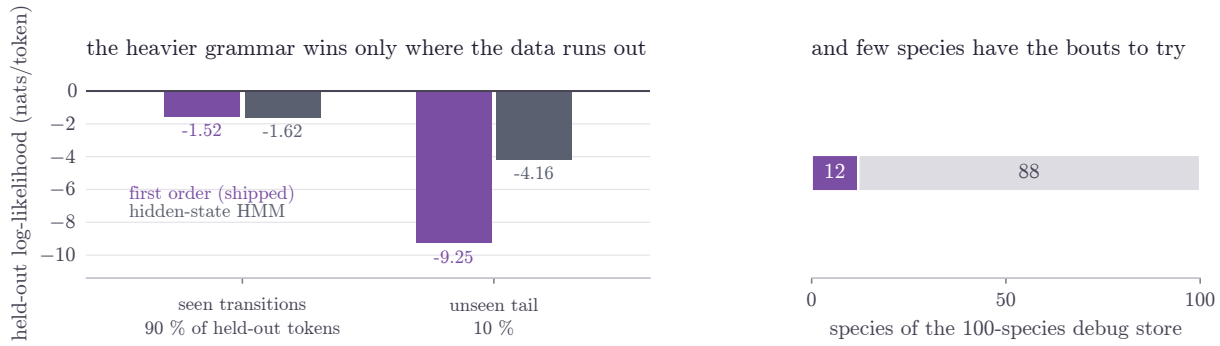
After the depth harvest the bake-off was re-run with qualifying species at 104, not 30. The HMM won again, at +0.4976 nats and 78 of 104 species, and this time against a partially repaired baseline: subtracting the continuation backoff's own measured gain on the same splits leaves the HMM ahead by about +0.37 nats per token. The debug-era explanation cannot account for a residual that size, so something closer to genuine hidden structure is scoring better now that the data can support it. It has *not* been adopted: the shipped inventory grammar is raw first-order counts, there is no HMM-state grammar writer, and adopting one changes what the fit emits, what the application improvises from and what song typing aligns against. That is recorded as the next syntax task with the number to beat.

8.9 Song-level structure: an edit-distance kernel against bag-of-types

Whole bouts as points, as (token, gap) sequences under a weighted edit-distance kernel with the rhythm term priced inside the dynamic program, judged against plain bag-of-types over the same bouts. It fails both pre-registered gates on both debug stores: species nearest-neighbour lift +0.877 against bag-of-types' +0.924 on the 37-species store and +0.050 against +0.081 on the 100-species store, with recordist lift +0.264 and +0.087 against a < 0.05 bar. Three measured details survive the failure. Within-species song

Table 6: Held-out cross-entropy by Markov order, real corpus against its within-species shuffle. 34 696 bouts, 327 164 tokens, 2 421 species; 217 species clear the 20-bout bar. The gate. Median gain ≥ 0.05 bits/token at order $1 \to 2$, ≥ 0.05 more than its own shuffle, and share of species gaining $\geq$ control +20 points. Fails on the magnitude criterion and passes on the share criterion, at three seeds.

	real	control
held-out cross-entropy, $k = 0$ (unigram)	3.284	2.879
$k = 1$	**2.301**	3.634
$k = 2$	2.546	4.868
$k = 3$	2.866	5.313
median gain, order $1 \to 2$ (bits/token)	**−0.091**	−0.912
median gain, order $2 \to 3$	−0.108	−0.187
share of species gaining at $1 \to 2$	**25.3 %**	4.1 %



Left: first order is the better model on the 90 % of held-out tokens whose transition was seen in training, and the hidden-state model's entire advantage is the unseen tail. That is a property of smoothing shape, not of recovered hidden structure, so the smoothing was repaired instead: a Kneser-Ney continuation backoff gains 1.90 and 2.32 nats per token on unseen-context transitions across two stores. Right: twelve species reach 20 bouts after run-collapse, which is the syntax threshold. Counted before the depth harvest; the production re-count is 217 of 2 421 species. Both panels are debug-scale.

Figure 16: Left: held-out likelihood split by whether the transition was seen in training. The heavier grammar's whole advantage lives where the data runs out. Right: species of the 100-species debug store with enough bouts to attempt song-level structure, counted before the depth harvest. Both panels are debug-scale.

structure exists (all 18 deep species separate their own recordings) and lives entirely in the token marginals, so order and rhythm add nothing measurable at current depth. The rhythm term actively *costs* species lift. And the bar for any learned song encoder is bag-of-types, not the edit kernel.

8.10 Model-free sequence statistics against their own shuffle

Three statistics that fit nothing, run on the production corpus against the same within-species permutation control. Conditional entropy of the next token given one previous is 0.941 bits against the shuffle's 3.343 and not a subtle effect. The Menzerath–Altmann exponent's median is +0.019 where the law predicts negative, with 45.7 % of species negative against the control's 8.5 %: a weak effect in the law's direction, not the law, and not the shape reported for geladas or penguins [19, 28]. The control's own exponent is consistently positive, which is a sampling artefact instead of a finding. A mean over more draws sits closer to the species mean and its logarithm rises with the count. That artefact is exactly why the control is run. Normalised compression distance [12] is refuted outright, because the control separates species *better* than the real corpus does, 0.038 against 0.018. Compression distance here reads alphabet and marginals, which the shuffle preserves, and not order. No learned song encoder should be measured against it.

One arm from this set was retracted by the project itself and is recorded as a design error, not as a result. Comparing the $k = 1 \to k = 2$ entropy drop between real and control is not a valid test when the two curves start 2.4 bits apart, and past $k = 3$ the control's plug-in curve falls *below* the real one because a plug-in estimator reports near-unique contexts as determined. Both effects are bias. The verdict function now refuses that comparison and not returning the answer it would have returned, and the unbiased instrument (held-out cross-entropy) is Table 6.

8.11 Detector configurations against a permutation control

Three configurations of the song detector were run against the same corpus-wide shuffle control, whose yield bar was 0.2 % of candidate bouts.

Repeat classes pass on both halves of their gate and are adopted. The control is unchanged to the bout, because shuffling a species' tokens destroys runs, so the control finds no repeat classes at all. The sharpest evidence available that the ones found are real.

The two refuted arms are instructive in the same

Table 7: Song-detector arms on the production store (2 421 species, 34 696 bouts, 10 550 candidate bouts of three tokens or more). Only the first passes. The control yield is the share of candidate bouts a within-species shuffle places, and the pre-registered bar is 0.2 %.

arm	species with a song	song types	placed bouts	control yield	verdict
shipped, motifs only	48	96	1 052	0.170 %	baseline
+ **repeat classes**	**98**	**356**	**3 741**	**0.170 %**	adopted
per-species recurrence bar	99–202	452–636	,	0.266–0.442 %	refuted
per-species bout rule	105	449	,	0.628 %	refuted

way. Letting each species set its own recurrence bar on the evidence of its own shuffled tokens let 2 398 of 2 421 species take the loosest bar, because a species with eight candidate bouts has a control that cannot place a bout at any size: "the control found nothing" meant "the control had nothing to work with". Adding a power requirement repairs the mechanism but not the verdict. No candidate threshold in $\{10, 15, 20, 30, 50\}$ returns the control to 0.2 %, and the rows that come closest gain almost nothing over the 98 species repeat classes already reach. A per-species bout rule at five times the species' own median gap reaches 105 species at nearly four times the chance yield, because splitting a fast singer's bouts finer produces more short strings and short strings are what chance can chain. Both were removed with their flags: a refuted policy nobody can reach is surface, not machinery, and the tables are what survive of them.

8.12 Four conclusions from the internal comparisons

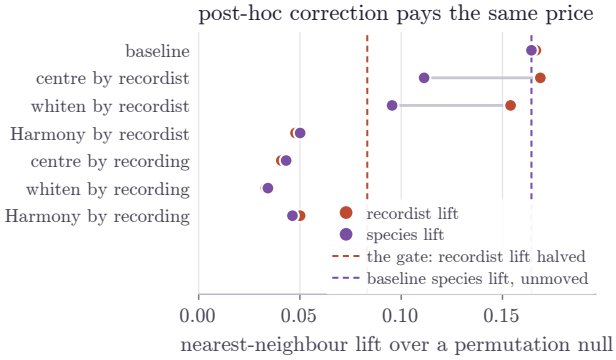
The choice of *space* matters most and is settled: the learned embedding beats the hand-weighted vector by 6 to 13 points of species placement, replicated. The choice of *clustering* matters exactly as much as the region budget binds, and no more. The choice of *objective* is unresolved and is the project's central open problem: the best available ruler is the worst available optimiser. And the choice of *starting point* does not matter, which is a genuinely useful negative because it eliminates one of the two remaining explanations for the fit's failure.

9 Our methods against other approaches

9.1 The limits of a fair comparison

The external comparisons in this section are all read-only tests on a fixed store, through one estimator. Nearest-neighbour lift over a label permutation null, blocked by recording, with no training and no new data. That makes them directly comparable to every lift this project has published. Three limits apply and are stated before the numbers, because they change what the numbers mean.

First, **the representation is per recording, not per syllable**. Perch's point is the mean of a recording's first six 5-second window embeddings; BirdNET's

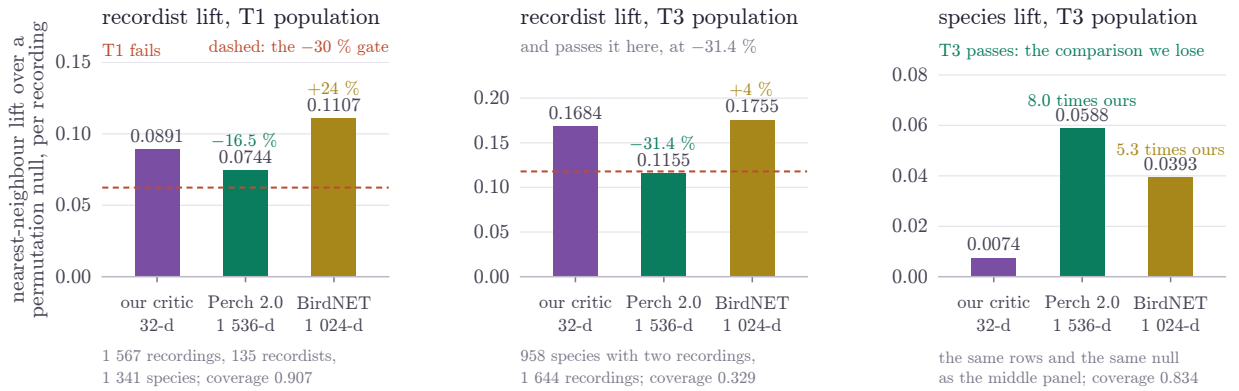


The gate asked for the recordist lift halved with the species lift not falling at all: the clay dot left of the clay rule and the violet dot still on the violet one. No arm does both. Subtracting each recordist's mean leaves the recordist lift higher than it started while costing a third of the species structure, so the confound is not a location shift.

Figure 17: Post-hoc batch correction of the production embedding reaches the confound and pays exactly the price a gradient-reversed adversarial head paid. The gate asked for the recordist lift halved with species lift not falling at all; no arm lands in it. Two unrelated mechanisms landing on the same exchange rate is evidence about the data, not about either mechanism, and the fact that subtracting each recordist's mean leaves the recordist lift *higher* than it started says the confound is not a location shift.

is the analogous average of its 3-second windows; the critic's is the mean of that recording's syllable embeddings. That form was chosen because two points from the same recording cannot exist in it, so same-recording leakage is impossible by construction, not by exclusion. But the critic's embedding is a *syllable* space being averaged over a recording, which is not the job it was fitted for. So a species-retrieval result here says that an external space is the better *per-recording* descriptor. **It does not say that Perch or BirdNET beats our critic syllable for syllable. That comparison has never been run.** The reason it has not is recorded: a per-syllable Perch pass over 340 079 syllables is about 3.3 hours on this machine's CPU, the installed ONNX runtime does not match the installed CUDA version, and the GPU provider fails to load and falls back to CPU silently at no measured speed difference.

Second, **the populations differ between the two draws and the difference is the result** and not noise to be averaged away. One draw was built for the recordist question, one for the species question, and they disagree about the recordist verdict. The tie-break was fixed on coverage before the disagreement



Read-only tests on a fixed store, one estimator, no training and no new data. The representation is per recording in every row, so same-recording leakage is impossible by construction - which also means a species result here says an external space is the better per-recording descriptor, and does not say Perch beats our critic syllable for syllable, a comparison that has never been run. T1's verdict rests on a coverage tie-break: it passes at −31.4 % on the T3 population and fails at −16.5 % on its own, whose recordist estimate covers 0.907 of rows against 0.329, so the instrument built for the question wins. The finding that holds in both draws is the one to carry: the recordist lift is twice the species lift in Perch's space, 4.5 times in BirdNET's and 23 times in ours. Every space tested is confounded in the same direction and they differ only in degree.

Figure 18: Three spaces, two populations. Recordist and species nearest-neighbour lift over a permutation null, per recording, for our 32-dimensional critic and two archive-scale external models. Perch 2.0 (trained on the same archives at roughly a hundred times our scale and never asked to tell our resynthesis from a bird) still carries five sixths of the recordist structure ours does, and BirdNET carries more than ours in both populations. If archive-scale training conferred recorder invariance the two largest models trained on these archives would both have it and they would agree; one is better than ours and one is worse, which locates the property in the archives. The right panel is the comparison we lose, and lose clearly. Hue carries model identity here and not a verdict.

appeared.

Third, **the audio cache constrains what can be asked at all**. It holds at most two recordings per species, 2 616 species with exactly two, 1 221 with one, none with three, because the iNaturalist ingest took no more and the xeno-canto audio was deleted at harvest by design. No species group of eight recordings exists on any audio this machine holds, so a species-stratified draw returns zero recordings and no selection rule can produce one. Anything needing several recordings of one species needs audio re-fetched first.

9.2 Recording-condition invariance: three spaces, two populations

T1 fails, and it is the fourth independent closure of the same road. Perch 2.0 is trained on the same archives at roughly a hundred times our scale, never asked to tell our resynthesis from a bird, and therefore structurally incapable of having learned the axis that splits our types in two. Still carries five sixths of the recordist structure our own critic carries. The gate asked for −30 % and measured −16.5 %. The prediction was written in advance and is worth quoting, not softening: *if a model trained on the same archives at a hundred times the scale is no more recorder-invariant than ours, the confound lives in the archives, not in our critic.* It does.

BirdNET fails harder, and the disagreement between the two external models is stronger evidence than either alone. BirdNET's recordist lift is *above* our critic's in both populations (+24 % and +4 %). Perch reads −16.5 % and BirdNET +24 % on the same rows. If archive-scale training conferred recorder invariance, the two largest models trained on these archives would both have it and they would agree. One is better than ours and one is worse, which locates the property in the archives and not in any model's loss. The same conclusion the batch-effect analysis reached from the inside, now reached from outside with two independent instruments.

T3 passes decisively, within its stated scope. Perch retrieves a species' other recording 8 times better than our critic does (0.0588 against 0.0074) and BirdNET 5.3 times better (0.0393), on the same rows against the same null. The direction replicates across both external models, so "an external space is the better per-recording species descriptor" holds. It licenses Perch as the per-recording batch descriptor for any future invariance work, and it weakens the case for building further external ports to answer T1, which is answered three ways.

The finding that holds in both draws is the one to carry. In Perch's space the recordist lift is still twice the species lift; in BirdNET's it is 4.5 times; in ours it is 23 times. Every space tested is confounded in the same direction and they differ only in degree. A model trained on these archives at a hundred times our scale, which has never seen a resynthesis, still encodes *who recorded it* more strongly than *which bird it is*.

Two things are explicitly not settled. T1's own verdict rests on a coverage tie-break: it passes at −31.4 % on the T3 population and fails at −16.5 % on the T1 population, and the T1 population's recordist estimate covers 0.907 of rows against the other's 0.329, so the instrument built for the question wins and the verdict is fail. The novelty comparison, Perch against our critic's novelty ROC of 0.846, was never run. Retiring the critic's species head was made conditional on T1 *and* T3 landing; T3 landed and T1 did not, so it stays.

Table 8: Recordist and species nearest-neighbour lift over a permutation null, for our 32-dimensional critic and two archive-scale external models, on two populations. Per recording in every row; same rows and same null within each population. *Coverage* is the share of rows for which the estimate is defined. T1's gate asked for a 30 % relative reduction in recordist lift; T3's asked for a species-retrieval improvement.

space	dims	recordist lift	species lift
T1 population: 1 567 recordings, 135 recordists, 1 341 species; recordist coverage 0.907			
critic	32	0.0891	not estimable
Perch 2.0	1 536	**0.0744** (−16.5 %)	not estimable
BirdNET	1 024	0.1107 (+24 %)	not estimable
T3 population: 958 species with two recordings each, 1 644 recordings, 760 recordists; species coverage 0.834, recordist coverage 0.329			
critic	32	0.1684	0.0074
Perch 2.0	1 536	**0.1155**	**0.0588**
BirdNET	1 024	0.1755	0.0393

9.3 Post-hoc batch correction against training-time invariance

The same confound was attacked with the genomics batch-effect toolkit, which has the property that its strongest members are post-hoc: they correct an embedding that already exists. 20 000 syllables of the production embedding, 885 species, 1 001 recordists, 3 265 recordings, every arm scored by the same estimator against its own null.

Three findings, in increasing order of consequence. The confound is not a location shift: subtracting each recordist's mean leaves the recordist lift higher than it started while costing a third of the species structure, which is a mechanistic explanation for why two training-time attempts failed and one that was never available before, because both of those attacked the confound during training where nothing can be inspected. Harmony [38] reaches it and pays exactly the price a gradient-reversed adversarial head paid: −71 % recordist, −70 % species. Two unrelated mechanisms landing on the same exchange rate is evidence about the data, not about either mechanism. And correcting by recording removes both because the labels are nearly nested. A recording normally holds one bird of one species, which is a statement about how the corpus was collected.

Against that, the data road works and its effect size shrank when re-tested. Figure 19 shows both scales. Merging two genuinely different recording domains, iNaturalist soundscape and xeno-canto focal, bought +14.4 points of cross-domain species retrieval on the 12 species present in both stores, with a 3-species holdout. The pre-registered bigger re-test on 63 both-domain species with 16 held out took that to +3.6. A cross-domain contrastive mechanism on top reached +13.4 at the small scale and +1.7 against a +10 gate at the larger one, so the mechanism's boost was small-sample structure. The uncomfortable part is recorded plainly: the merge's *own* win shrank too, and its held-out realism AUC reads below the single-domain arm's here (0.976 against 0.995) where the first test gained 0.012. Every number from the first test should be treated as small-sample in both directions. What survives the bigger test is the ordering, which is the load-bearing claim: more domains per species helps, and the mechanism adds approximately nothing on top of the data.

9.4 Positioning

Table 10 places the system beside the approaches it is most likely to be confused with. A cell reads "not documented" where the cited sources do not answer it; the column entries for this work are properties of the shipped artefact, not aspirations.

The comparison that matters is not on accuracy, because the systems answer different questions. The external models are better at "which bird is this", by a margin that has been measured here at 5 to 8 times on per-recording species retrieval, and this work should be read as taking that for granted, not competing with it. What this work provides that none of them does is a *generative, physically parameterised* account of a syllable that is fitted to field audio at archive scale, together with an explicit statement of the evidence behind every claim built on it.

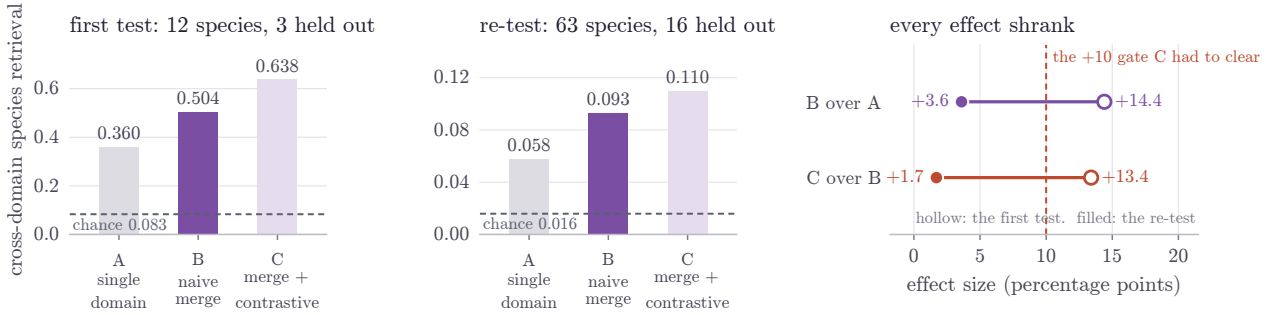
10 Discussion

Four of this paper's results generalise past the system that produced them.

A gated proxy can be anti-correlated with the thing you actually care about, and pass its own gate the whole time. Every fit, sweep and ship decision in this pipeline was judged on one scalar. Over 1 200 types that scalar is significantly correlated with audible distortion in the *wrong* direction, and the first adoption it gated that could be re-judged cheaply turned out to have degraded the sound by 4 dB while gaining a third of a cosine point. Neither fact was visible from inside the gate, because a gate cannot audit its own metric. What made it visible was building a second instrument that owed nothing to the first, and then spending the effort to re-judge a decision already shipped. The second adoption re-judged the same way then *survived* it, which is the part worth carrying: a bad proxy does not corrupt every decision it gates, so the audit cannot be skipped in either direction. One instance is not a rule, and one counter-instance is not an acquittal — each shipped gate has to be re-judged

Table 9: Post-hoc batch correction of the production embedding. 20 000 syllables, 885 species, 1 001 recordists, 3 265 recordings. The gate asked for the recordist lift halved with species lift not falling at all. No arm passes: every arm that halves one halves the other.

arm	species lift	recordist lift	ratio
baseline	0.1643	0.1665	0.99
centre by recordist	0.1113	0.1687	0.66
whiten by recordist	0.0955	0.1540	0.62
Harmony by recordist	0.0501	0.0478	1.05
centre by recording	0.0432	0.0408	1.06
whiten by recording	0.0342	0.0331	1.03
Harmony by recording	0.0463	0.0501	0.92



Merging two genuinely different recording domains - iNaturalist soundscape and xeno-canto focal - is the data road; the contrastive mechanism on top is the method road. Hollow marker in the right panel: the first test. Filled: the pre-registered re-test. The mechanism's boost was small-sample structure and the data road's was not, though it survives at a quarter of the size. The merge's own held-out realism AUC reads below the single-domain arm's in the re-test (0.976 against 0.995) where the first test gained 0.012. Both panels are debug-scale and absolute levels are comparable only within a panel.

Figure 19: Cross-domain species retrieval for three training arms at two scales. The left two panels are the two scales; both are debug-scale, and absolute levels are comparable only arm against arm inside each panel. Right: the two effect sizes at both scales against the +10 point adoption gate, which is the panel that carries the verdict — the pre-registered re-test shrank every effect and no arm reaches the gate, while the ordering survived.

on its own. Any project with one number in the loop — a reward model, a learned critic, a perceptual loss — can be in this position and cannot find out from the inside.

Recording-condition confounding is a property of the public bioacoustic archives, not of any one model's loss. It has now been attacked from four independent directions. A gradient-reversed adversarial head, a species-factored version of the same, the post-hoc genomics batch-effect toolkit including its strongest member, and two archive-scale external embeddings, and every route either fails to reach it or pays for it one-for-one in species structure. The two external models disagree with each other by more than either differs from ours, which is the decisive form of the argument: if archive-scale training conferred the property, both would have it and they would agree. Since most species in these archives are carried by a handful of recordists, no method that sees only an embedding can separate labels the sample never separated. The prescription is a data action. More recordists, and more genuinely different recording domains, per species, and it now has two independent derivations, one from domain merging and one from batch correction.

An excellent discriminator is not an objective. A critic that separates real from synthetic at AUC 0.990 on entirely held-out species is monotonically *harmful* when its logit is added to the fit objective. Separating two distributions is not the same as having a gradient that points at the nearer one. Retaining a guardrail metric the learned term cannot game is what made this visible, not a mystery, and that arrangement generalises to any learned-metric-in-the-loop setup.

Small samples flatter the arm you hoped for, in both directions. The project measured this twice, at similar cost. A cross-domain contrastive mechanism worth +13.4 points on 12 species with a 3-species hold-out was worth +1.7 on 63 species with 16 held out; in the same re-test the arm it was compared against also shrank, and one of its metrics reversed sign. An objective change winning 7 of 10 types won 25 of 80. And a correlation that was "indistinguishable from zero" at $n = 200$ was significant at $p = 3 \times 10^{-12}$ at $n = 1\,200$, with the sign it had always had — the small sample hid a real effect by leaving it inside the error bars, which is the same failure wearing the opposite face. In each case the corrective was a pre-registered re-test at scale and not an argument about whether the first result was noise.

For anyone building a comparable system, the ordering the evidence supports is: fix the representation before the clustering, fix the corpus before the loss, and do not spend anything on the search's starting point. And write the gate down first, because three of this project's most useful findings are gates that failed in ways that identified the real constraint. The label-assignment arm, whose failure showed that seeded k-means is nearly as good as knowing the answer where

Table 10: Approaches beside each other. External rows are read from the cited sources; "not documented" means the source does not answer the cell. *Interpretable parameters* means a reader can name the quantity that produced a given sound.

approach	what it is	sings	interpretable parameters	runs as a static asset set	labels its claims
BirdNET [35]	species classifier on spectrograms	no	no	no	not documented
Perch 2.0 [73]	general bioacoustic embedding, ∼15k classes	no	no	no	no
AVES [29]	self-supervised animal audio encoder	no	no	no	no
BirdMAE [59]	masked autoencoder for bird audio	no	no	no	no
NatureLM-audio [64]	audio-language model for bioacoustics	no	no	no	no
Sainburg et al. [66]	projected latent space across repertoires	no	latent and not physiological	not documented	no
Goffinet et al. [25]	learned low-dimensional vocal features	no	no	not documented	no
generative audio models	waveform generators	yes	no	no	no
Lyrebird	physical syrinx + a corpus with no audio in it	**yes**	**yes** (pressure, tension, tract geometry	**yes**) no audio retained, no weights shipped	**yes**. Five levels

the cluster budget does not bind; the inverse encoder, whose failure eliminated the starting point as an explanation; and the second-order syntax arm, whose failure converted a method question into a countable data question.

11 Limitations

Scale of the evidence. Many comparative numbers here come from a 100-species debug subset, a 37-species store, or, for the domain-merge arms, 12 or 63 species. That is a deliberate operating rule. We adopted it after a full-corpus build exhausted the machine's memory, and every occurrence is marked. The production-scale claims are the corpus counts, the fit medians, the sequence statistics and the built asset set. Most arm-against-arm comparisons are not production-scale.

Self-reported metrics. Under label assignment, cluster purity and type coverage are identities and not measurements. The shipped manifest says so. Type retrieval survives because nothing forces it, so type retrieval is the figure we quote.

No audible evaluation. We ran no listening test, no mean-opinion score and no playback experiment. The fidelity claims rest on mel-cepstral distortion and multi-resolution spectral distance against cached audio, over 1 200 types. The objective that produced those fits is positively correlated with both metrics, meaning it ranks audible fidelity *backwards*. Two of the four cosine-gated adoptions have since been re-judged on the bench: respiration lost and the two-voice fit held. The other two still owe that re-judgment and both need a refit to get it.

Unresolved internal conflicts. T1's verdict rests on a coverage tie-break between two populations that disagree. The fixed-waveguide tract improves everything in isolation, fails across 22 species on one store, and wins under learned typing on another. The three-dimensional projection's distance faithfulness fell from 0.903 to 0.747 when we removed the synthetic anchors. On-screen distance now means less than it did, although nothing about the regions or the matching uses those positions.

Production-scale re-measurement, and what it cost. The map-agreement decomposition had only been measured on a debug build, and its own note asked for a production re-run before any production decision. The production build now carries those figures, and they are worse than the debug build predicted. Region agreement is **0.4909** against 0.6316. Tie-adjusted agreement is **0.7056** against 0.8136. Species-level agreement is **0.5619** against 0.6831. Each of the three loses 12 to 14 points.

The direction is explicable. Agreement is a nearest-centroid lookup, and this build holds 14 003 knots where the debug build held 4 298. A held-out syllable therefore has three times as many near neighbours available to be confused with the right one.

The consequence matters for the interface's central gesture. On the shipped production build, a live syllable routed through the descriptor map reaches the region the embedding itself would choose **less than half the time**, and the same species just over half. The tie-adjusted figure shows that most of the disagreement is a near-miss between adjacent knots, which is the difference between a blurred answer and a wrong one. It does not make the raw figure larger, and the raw figure is the one to hold. Taken with the 0.602 retrieval ceiling of §8, the bound on "sing and the right region lights" at production scale is lower than any debug number in this paper.

Unrun. Five things carry a gate in the project's plan

document and have not been run:

- a per-syllable comparison against the external models;
- a novelty comparison against those same models;
- the occurrence prior, which is blocked on coordinates the harvester never recorded;
- simulation-based inference and differentiable DSP for the fit;
- weak function labels from the archives' own behavioural annotations.

12 Two questions for comparative biology

This paper is a system, a resource and a ledger of refusals. It is not a contribution to ornithology, and the distinction is worth making because the corpus can carry two questions that would be. Both are specified with pre-registered gates in the project's plan document, and both need no new audio.

Both were subsequently run. The two paragraphs that follow were written before the runs and are kept as written; Section 12.1 reports what they answered, and the short version is that neither opened the door it was aimed at. The corpus is still not a contribution to ornithology, and it is now that by measurement, not by caution.

Is benchmark accuracy in this field inflated by recordist leakage? Standard evaluation splits recordings at random. Section 7 measures a nearest-neighbour lift of +0.488 for recordist against +0.196 for species, and Section 9 finds the two archive-scale external models carry it as well. If a species' recordings come disproportionately from one recordist, a random split leaks recorder identity across the train and test sides. The test is one comparison on a public corpus with public models, changing only the split: random against recordist-blocked. A large drop would mean the field should block its splits by recordist, which is a cheap change with a measurable cost. This result would not depend on any part of the pipeline described here being correct, which is what makes it the widest-reaching thing the project can offer.

Can anatomy be inferred from song at a scale dissection cannot reach? The fit estimates an effective tracheal length for 1 621 species. Tracheal length scales with body size, and several lineages carry documented elongation. Regressing the fitted length on published body mass, controlling for the species' own pitch, decides whether the parameter is anatomy or a fitting artefact. Three facts are known before that run and none of them is encouraging on its own: the length is a deterministic restatement of the fitted tract resonance, it is quantized to 39 values, and 24 % of species sit on its lower clamp. The single encouraging fact is that it correlates with the species' median f_0 at only -0.206, so it is not pitch in centimetres. If it survives, comparative work on song *production*, not song *sound* becomes possible, and that has not been available before at this scale. If it does not, the negative result is worth reporting on its own.

12.1 The answers

The two paragraphs above were written before either run. Both ran on 2026-08-04, against pre-registered gates that were not moved afterwards, and the paragraphs are left standing so that the prediction and the outcome can be read together. Table 11 is the full ledger with the command that reproduces each row. Ordered by what a reader should take from them:

Anatomy cannot be inferred from song at this scale, and the two available excuses are closed. The fitted tracheal length correlates with published body mass at a partial r of +0.194 controlling for the species' own pitch, over 1 187 unclamped species joined to AVONET. The gate asked for 0.4. The correlation is real and the direction is right — $p = 1.5 \times 10^{-11}$ — and it is less than half of what was required. Neither obvious defence survives measurement: a *perfect* predictor pushed through the same 39 quantization levels still reaches $r = +0.998$, so the grid is not the ceiling; and the correlation is flat as the per-species sample grows (+0.241, +0.263, +0.226, +0.299), so it is not thin sampling. The one encouraging fragment is that cranes sit 2 of 2 in the upper decile at a median 10.6 cm, on a sample too small to carry weight. **Comparative work on song production is closed until a tract estimate passes this gate**, and by the plan's own rule that also closes the phylogenetic-signal and source–filter questions that depend on it.

The pipeline recovers the textbook relationship, which is what makes the negative result above worth reading. Species median f_0 against body mass under phylogenetic generalised least squares over ten BirdTree posterior trees gives a standardised slope of -0.560 on deeply sampled species, against the -0.510 that Mikula et al. report over 5 085 passerines. A pipeline that failed both gates would simply be broken; one that reproduces a known result at the published effect size and still cannot find anatomy in its own fitted parameter has localised the failure.

Two method points travel with that number. **Pagel's λ must be fitted rather than assumed.** Forcing Brownian motion — which is what "PGLS" is often taken to mean — returned standardised slopes spanning -1.962 to $+0.001$ across ten trees drawn from one posterior, a range that crosses zero. That was not numerical: every pruned tree was exactly ultrametric and positive definite with condition numbers around 10^6–10^7. Fitting λ collapsed the range to $[-0.214, -0.207]$. And **the fitted λ is itself a measurement of our own noise**: it comes out at 0.000 on 651 species and 0.291 on the 76 most deeply sampled, and a λ of zero for song frequency cannot be true of

birds. Measurement error drives λ towards zero and attenuates a standardised slope by the same mechanism. The shortfall against -0.510 at full sample is ours and not the birds'.

We tested whether that is really a depth effect and the test failed, so we do not claim one. Turning the three points into a curve — seven sampling floors, each paired with a matched-n control subsample of the unfiltered species set — pre-registered three clauses, of which the second decided the reading: the controls must not reproduce the climb. They did. λ rises 0.106 to 0.319 across the depth arms, but the control arms spread 0.230 against a bar of half the depth spread, so the depth interpretation is withdrawn as pre-registered. A follow-up says why the test could not separate them, and it indicts the estimator rather than the data: drawing 20 control subsamples per size instead of one, maximum likelihood puts λ exactly at its lower boundary in 85 to 95 % of draws at $n \leq 123$ and in 0 % at $n = 1\,379$. At the sizes where the deep arms live, a single λ estimate is mostly a coin flip against zero, and the interval our curve reported spanned posterior trees rather than the sampling of species, so it read as zero-width next to arms differing by 0.1. Five of six depth arms do sit above the 95th percentile of the null band at their own n, which is why we think the effect is probably real — but that comparison was designed after seeing the pre-registered one fail, so it is a lead for a properly pre-registered test and not a result. What stands is narrower and still useful: at $n = 651$ a fitted λ of 0.000 is unremarkable, because a third of random draws of 651 species from this same data produce it.

A recordist-blocked split costs a probe over frozen foundation-model features 18.51 points of top-1 on Perch and 21.59 on BirdNET, and the gate is met on both. The population is **8 287** recordings over **884** species from **646** recordists, all uploaded to xeno-canto after both models' training snapshots. A linear probe over frozen Perch 2.0 features reaches 81.68 % under a random split and 63.16 % under a recordist-blocked one. On frozen BirdNET V2.4 features it is 79.08 % and 57.49 %. Blocking was worse in **20 of 20 paired folds** on both models, per-fold standard deviation under 3 points, and the two agree on the error ratio to two decimal places — 2.01× and 2.03× — while differing by 3 points in the raw figure. The pre-registered bar was more than 10 points reproduced on both models, so it is met.

The pilot said 9.68 points and failed the gate, and its population is why. It read a catalogue of exactly 6 000 recordings, which was its own 60-page search limit and not the archive: the same query holds 21 928. And it selected by *recordist*, a rule inherited from the T1b arm whose unit was the recordist, which left 62 scorable species out of 827 fetched. Run at 62 classes on this population the same procedure gives 12.75 points, so part of the pilot's shortfall was its label set and part was its sample.

The gate was written in the wrong unit, and the full run shows why more clearly than the pilot did. Across nested label sets of 62 to 884 species on the same recordings and folds, the drop climbs from 12.75 to 18.51 points while the error ratio *falls* from 2.97× to 2.01×. We had predicted the ratio would be flat. It is not, so there is no unit in which this effect is one number independent of the label set, and any figure quoted without its class count cannot be compared against. The bar should have been an error-rate ratio with the label-set size stated. It is not being moved after the fact.

The design also had to change before it could run at all: *both* candidate models train on xeno-canto, and Perch additionally on iNaturalist, so a top-1 comparison on archive audio cannot separate leakage from memorisation. Freezing the model and training only a linear probe puts identical contamination on both sides of the comparison, and post-cutoff audio removes it entirely. Because the discriminator is upload date and not archive provenance, this also makes BirdNET testable on an xeno-canto corpus, which the registered method had ruled out: V2.4 was released in June 2023, two years before the cutoff.

The recordist structure is visible to an outside model on recordings it has never seen. On that same post-cutoff population, Perch's nearest-neighbour recordist lift is +0.441 against a species lift of +0.750, both scored on identical rows at full coverage, both strong, both at the permutation floor. Section 7 reports the confound on our own store and Section 9 on archive-scale models evaluated over iNaturalist audio that Perch had trained on; this arm removes that objection. The confound is a property of the archives, not of any one model's memorisation.

Aggregating to one point per recording does not remove the confound, and the test that appears to say otherwise is selecting its own answer. Scored per recording rather than per syllable, species lift (+0.229) exceeds recordist lift (+0.196) on rows where both groupings clear the estimator's eight-member floor. That reads as an inversion and is not one. The same population scored at syllable level already puts species above recordist (+0.343 against +0.295), and aggregation lowers both by about a third together. What differs is the *population*: at recording level the eight-member floor means "species with at least eight recordings", which is 77 of 3 636 — and recordings per species correlates with distinct recordists per species at $r = +0.923$ (median 17 recordists for species above the floor, 2 below it). **The floor that makes the test possible is the floor that removes what it is testing for.** Section 7 already states the underlying fact in words; this is a number for it, and a warning about a test design.

Whether vocal learners have more complex song cannot be asked of this corpus, and that is a statement about our measures and not a null about birds. The contrast ran at full scale —

2 059 species, learner status from BirdTree's own oscine/suboscine column, PGLS over ten posterior trees, body mass and sampling effort as covariates — and every coefficient came back flat. It should not be reported as a null. Our repertoire proxy correlates with how much audio we hold for a species at $r = +0.852$, and the repeat-class count is zero for 97.7 % of species. One of three measures is degenerate and another is largely a count of our own collection effort, so a two-of-three gate cannot be cleared by what remains. The binding constraint is depth per species, at a median of 69 syllables.

Table 11: Phase-0 findings, most consequential first, with the gate set in advance and the command that reproduces each row. Every runner prints its own gate before its result and writes a JSON sidecar beside itself.

finding	the gate, set in advance	what it measured	how to reproduce
G1 — fitted tract length against body mass	partial $r > 0.4$ controlling $\log f_0$, over ≥ 800 unclamped species, and the known long-trachea clades in the upper decile	$+0.194$ over 1 187 species. A perfect predictor through the same 39 levels reaches $+0.998$, and r is flat in per-species sample size, so neither quantization nor sampling explains it	`analysis/g1_g2_avonet.py`; needs AVONET (figshare 16586228)
G2 — species median f_0 against body mass	negative slope, significant under PGLS, effect size inside the published range	standardised slope -0.560 on deeply sampled species against a published -0.510; λ fitted at 0.130 overall and 0.000 at shallow depth	`analysis/g2_pgls.py`; needs BirdTree Ericson Stage-2 trees
W0 — cost of a recordist-blocked split	> 10 points of top-1 drop, on a public corpus, on both models	**MET.** 18.51 points on Perch ($81.68 \rightarrow 63.16$) and 21.59 on BirdNET ($79.08 \rightarrow 57.49$), 20 of 20 folds each, $2.01\times$ and $2.03\times$ the errors, over 884 species and 646 recordists. The pilot's 9.68 was one model at pilot scale	`analysis/w0_full_fetch.py` then `w0_full.py`; the pilot is `w0_recordist_split.py`
W0b — how many recordists per species it takes	none; this was not registered, and it is reported as exploratory	30.82 points of drop at 2 recordists per species falling to 3.08 at ≥ 6, with label-set size and recordings per species held flat. The same recordists are findable in both models at a partial r of $+0.852$	`analysis/w0_confound_anatomy.py`
T1b — Perch's recordist lift on unseen audio	is the structure present at all on audio outside the training snapshot	recordist $+0.441$, species $+0.750$, matched rows, both strong	`analysis/w6_t1b_perch_postcutoff.py`; fetches its own audio
G3 — confound after aggregation	species lift must exceed recordist lift; state the margin	$+0.229$ against $+0.196$, but on 77 of 3 636 species, and recordings per species correlates with recordists per species at $+0.923$	`analysis/g3_recordist_aggregation.py`, ARM D
W2 — learners against non-learners	learners exceed on ≥ 2 of 3 measures after phylogenetic correction	refused. Repertoire against sampling effort $r = +0.852$; repeat classes zero on 97.7 % of species	`analysis/w2_learners.py`

One caveat applies to every row and is a consequence of the corpus, not of any method: xeno-canto hosts bats, frogs, grasshoppers and land mammals alongside birds, and a query without an explicit group tag returned 6 000 recordings of which none were birds. It raised no error. Any replication of the last two rows must carry `grp:birds`.

13 Conclusion

Lyrebird measures 378 224 syllables from wild recordings without keeping the recordings, fits a published physical syrinx model to 16 283 syllable types over 2 650 species, learns a measured grammar over those types, and displays the result as a map in which live sound lights the acoustically nearest regions, with an evidence label on every claim.

The results that survive scrutiny are narrower than the system. A first-order transition model over the learned tokens buys 0.98 bits per token held out while costing 0.75 on the shuffle, which establishes that the tokens are units. Repeat classes take the corpus from 48 to 98 species with a measured song while the shuffle control does not move. A second order is not supported, and 217 of 2 421 species have enough bouts to ask. Type retrieval tops out at 0.602, so for two syllables in five the nearest region is the wrong one. The objective that produced every fit ranks audible fidelity *inversely*, and one adoption it gated has been measured degrading the sound. And the embedding organises by recordist before it organises by bird. A property of the archives, established four independent ways, including from outside our own loss entirely.

Thirty-six of fifty-eight pre-registered gates failed. The complete ledger is Appendix A. Those refusals are the part of this work most likely to save someone else time, and they are the reason the positive results above can be stated without hedging.

Reproducibility and data availability

The pipeline, the browser application and the complete research record are at `https://github.com/sha5b/Lyrebird`, and the running application is at `https://lyrebird.variable.gallery`. No archived release DOI exists: the work is a moving record rather than a versioned artefact, and minting one would date a repository that is still being measured. The commit the numbers here were read from is the state of that repository on the date of this preprint. The harvest stores are not distributed: they contain no audio, but they were gathered under archive rate limits and are the only copies. Every one-off measurement in Sections 8 and 9 is a committed script whose pre-registered gate is the docstring at the top of the file, on the principle that a result whose gate cannot be re-read is not evidence. Two fixed artefacts the comparative arms depend on (a dumped clustering matrix and a local audio cache) sit outside the repository by design, so absolute numbers will move on regeneration while the arm-against-arm comparisons hold.

Ethics and licensing

All recordings come from xeno-canto and iNaturalist under their per-file licences. Recordings under a NoDerivatives licence are refused at harvest, because every downstream product is a derived measurement, and recordings with no recorded licence are refused, because absence of a licence is not permission. Ninety point three per cent of the production store's 10 184 recordings pass, and all 10 184 carry an attributed recordist. No audio is redistributed and none reaches the browser. External model weights are used offline only, under their own licences, and their outputs are recorded as metadata that no part of the inventory may read. This is a non-commercial research project.

Acknowledgements

This work owes its existence to people who released their results, their recordings and their code for others to build on. The physical syrinx model is not ours: it is taken from the published normal-form work and used term for term, and the fits reported here are only possible because that literature stated its equations and its parameter ranges plainly enough to be reimplemented. The corpus exists because tens of thousands of recordists uploaded field audio to xeno-canto and iNaturalist under licences that permit derived measurement, and because both archives publish it with attribution intact. The external opinions that closed this paper's central negative result four ways come from Perch, BirdNET, AVES, BirdMAE and NatureLM-audio, all of which were released with usable weights; a self-reported embedding quality number is worth much less than one an outside model can contradict, and that contradiction was only available because those models are open. The methods borrowed from other fields — Harmony from genomics, MixStyle and gradient reversal from domain adaptation, HDBSCAN* and UMAP and random Fourier features from the clustering and kernel literature — were likewise all reimplementable from open descriptions and open code.

Funding

None. This is independent work by an artist, made at the boundary of art and science, with no grant, institutional or commercial support of any kind.

References

[1] Ana Amador and Gabriel B. Mindlin. Beyond harmonic sounds in a simple model for birdsong production. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 18(4):043123, 2008. doi:

10.1063/1.3041023. URL https://doi.org/10.1063/1.3041023.

[2] Ana Amador and Gabriel B. Mindlin. Synthetic birdsongs as a tool to induce, and listen to, replay activity in sleeping birds. *Frontiers in Neuroscience*, 15:647978, 2021. doi: 10.3389/fnins.2021.647978. URL https://doi.org/10.3389/fnins.2021.647978. Review article; the noise-on-tension result it reproduces originates in amador2013.

[3] Ana Amador, Yonatan Sanz Perl, Gabriel B. Mindlin, and Daniel Margoliash. Elemental gesture dynamics are encoded by song premotor cortical neurons. *Nature*, 495(7439):59–64, 2013. doi: 10.1038/nature11967. URL https://doi.org/10.1038/nature11967.

[4] Ezequiel M. Arneodo, Yonatan Sanz Perl, Franz Goller, and Gabriel B. Mindlin. Prosthetic avian vocal organ controlled by a freely behaving bird based on a low dimensional model of the biomechanical periphery. *PLOS Computational Biology*, 8(6):e1002546, 2012. doi: 10.1371/journal.pcbi.1002546. URL https://doi.org/10.1371/journal.pcbi.1002546.

[5] Gabriël J. L. Beckers, Roderick A. Suthers, and Carel ten Cate. Pure-tone birdsong by resonance filtering of harmonic overtones. *Proceedings of the National Academy of Sciences*, 100(12):7372–7376, 2003. doi: 10.1073/pnas.1232227100. URL https://doi.org/10.1073/pnas.1232227100. Species are doves (*Streptopelia*), not passerines.

[6] Robert C. Berwick, Kazuo Okanoya, Gabriël J. L. Beckers, and Johan J. Bolhuis. Songs to syntax: the linguistics of birdsong. *Trends in Cognitive Sciences*, 15(3):113–121, 2011. doi: 10.1016/j.tics.2011.01.002. URL https://doi.org/10.1016/j.tics.2011.01.002.

[7] Roberto Bistel, Ana Amador, and Gabriel B. Mindlin. Response of wild songbirds to songs synthesized with a low-dimensional model. *Physical Review E*, 109(5):054410, 2024. doi: 10.1103/PhysRevE.109.054410. URL https://doi.org/10.1103/PhysRevE.109.054410.

[8] Santiago Boari, Yonatan Sanz Perl, Ana Amador, Daniel Margoliash, and Gabriel B. Mindlin. Automatic reconstruction of physiological gestures used in a model of birdsong production. *Journal of Neurophysiology*, 114(5):2912–2922, 2015. doi: 10.1152/jn.00385.2015. URL https://doi.org/10.1152/jn.00385.2015.

[9] Ricardo J. G. B. Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining (PAKDD 2013)*, volume 7819 of *Lecture Notes in Computer Science*, pages 160–172. Springer, 2013. doi: 10.1007/978-3-642-37456-2_14. URL https://doi.org/10.1007/978-3-642-37456-2_14.

[10] Mustafa Chasmai, Alexander Shepard, Subhransu Maji, and Grant Van Horn. The iNaturalist sounds dataset. In *Advances in Neural Information Processing Systems 37 (NeurIPS), Datasets and Benchmarks Track*, 2024. URL https://arxiv.org/abs/2506.00343. Preprint posted 2025, after the NeurIPS 2024 presentation.

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proc. 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607, 2020. URL http://proceedings.mlr.press/v119/chen20j.html.

[12] Rudi Cilibrasi and Paul M. B. Vitányi. Clustering by compression. *IEEE Transactions on Information Theory*, 51(4):1523–1545, 2005. doi: 10.1109/TIT.2005.844059. URL https://doi.org/10.1109/TIT.2005.844059. Preprint arXiv:cs/0312044 (2003). The normalised compression distance.

[13] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020. doi: 10.1073/pnas.1912789117. URL https://doi.org/10.1073/pnas.1912789117.

[14] Anastasia H. Dalziell and Robert D. Magrath. Fooling the experts: accurate vocal mimicry in the song of the superb lyrebird, *menura novaehollandiae*. *Animal Behaviour*, 83(6):1401–1410, 2012. doi: 10.1016/j.anbehav.2012.03.009. URL https://doi.org/10.1016/j.anbehav.2012.03.009. Shrike-thrushes reacted as strongly to mimetic song as to their own when songs were presented *alone*; approach dropped when the mimicry was embedded in lyrebird song sequences.

[15] Coen P. H. Elemans, Andrew F. Mead, Lawrence C. Rome, and Franz Goller. Superfast vocal muscles control song production in songbirds. *PLOS ONE*, 3(7):e2581, 2008. doi: 10.1371/journal.pone.0002581. URL https://doi.org/10.1371/journal.pone.0002581.

[16] Coen P. H. Elemans, Riccardo Zaccarelli, and Hanspeter Herzel. Biomechanics and control of vocalization in a non-songbird. *Journal of the Royal Society Interface*, 5(24):691–703, 2008. doi: 10.1098/rsif.2007.1237. URL https://doi.org/10.1098/rsif.2007.1237. Cited in the repository as 2009, vol. 6, pp. 491–506; all three are wrong.

[17] Jesse Engel, Lamtharn Hantrakul, Chenjie Gu, and Adam Roberts. DDSP: Differentiable digital signal processing. In *Proc. International Conference on Learning Representations (ICLR)*, 2020. URL https://arxiv.org/abs/2001.04643.

[18] Facundo Fainstein, Franz Goller, and Gabriel B. Mindlin. Nonlinear biomechanical resonances in birdsong, 2025. URL https://arxiv.org/abs/2509.11019. Preprint, v2. The v1 title was "Birds breathe and sing at resonances of the biomechanics"; cite the version.

[19] Livio Favaro, Marco Gamba, Eleonora Cresta, Elena Fumagalli, Francesca Bandoli, Cristina Pilenga, Valentina Isaja, Nicolas Mathevon, and David Reby. Do penguins' vocal sequences conform to linguistic laws? *Biology Letters*, 16(2):20190589, 2020. doi: 10.1098/rsbl.2019.0589. URL https://doi.org/10.1098/rsbl.2019.0589. African penguin. Menzerath–Altmann holds, but by shifting sequence composition toward the shortest syllable type rather than by abbreviating all syllable types as in gustison2016.

[20] Michale S. Fee, Boris Shraiman, Bijan Pesaran, and Partha P. Mitra. The role of nonlinear dynamics of the syrinx in the vocalizations of a songbird. *Nature*, 395(6697):67–71, 1998. doi: 10.1038/25725. URL https://doi.org/10.1038/25725.

[21] Neville H. Fletcher, Tobias Riede, and Roderick A. Suthers. Model for vocalization by a bird with distensible vocal cavity and open beak. *Journal of the Acoustical Society of America*, 119(2):1005–1011, 2006. doi: 10.1121/1.2159434. URL https://doi.org/10.1121/1.2159434.

[22] Todd M. Freeberg, Jeffrey R. Lucas, and Barbara Clucas. Variation in chick-a-dee calls of a carolina chickadee population, *poecile carolinensis*: identity and redundancy within note types. *Journal of the Acoustical Society of America*, 113(4):2127–2136, 2003. doi: 10.1121/1.1559175. URL https://doi.org/10.1121/1.1559175. Carolina chickadee, and about sex/microgeographic identity — not the A-B-C-D syntax and not D-count scaling.

[23] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016. URL https://jmlr.org/papers/v17/15-239.html. The gradient-reversal layer. No DOI (JMLR). Note that JMLR's own exported BibTeX truncates the seventh author to "Mario March"; the correct name is Marchand.

[24] Tim Gardner, Guillermo Cecchi, Marcelo Magnasco, Rodrigo Laje, and Gabriel B. Mindlin. Simple motor gestures for birdsongs. *Physical Review Letters*, 87(20):208101, 2001. doi: 10.1103/PhysRevLett.87.208101. URL https://doi.org/10.1103/PhysRevLett.87.208101.

[25] Jack Goffinet, Samuel Brudner, Richard Mooney, and John Pearson. Low-dimensional learned feature spaces quantify individual and group differences in vocal repertoires. *eLife*, 10:e67855, 2021. doi: 10.7554/eLife.67855. URL https://doi.org/10.7554/eLife.67855.

[26] Franz Goller and Roderick A. Suthers. Role of syringeal muscles in gating airflow and sound production in singing brown thrashers. *Journal of Neurophysiology*, 75(2):867–876, 1996. doi: 10.1152/jn.1996.75.2.867. URL https://doi.org/10.1152/jn.1996.75.2.867.

[27] Michael Griesser. Referential calls signal predator behavior in a group-living bird species. *Current Biology*, 18(1):69–73, 2008. doi: 10.1016/j.cub.2007.11.069. URL https://doi.org/10.1016/j.cub.2007.11.069.

[28] Morgan L. Gustison, Stuart Semple, Ramon Ferrer-i Cancho, and Thore J. Bergman. Gelada vocal sequences follow Menzerath's linguistic law. *Proceedings of the National Academy of Sciences*, 113(19):E2750–E2758, 2016. doi: 10.1073/pnas.1522072113. URL https://doi.org/10.1073/pnas.1522072113.

[29] Masato Hagiwara. AVES: Animal vocalization encoder based on self-supervision. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023. URL https://arxiv.org/abs/2210.14493.

[30] Masato Hagiwara, Benjamin Hoffman, Jen-Yu Liu, Maddie Cusimano, Felix Effenberger, and Katie Zacarian. BEANS: The benchmark of animal sounds. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10096686. URL https://doi.org/10.1109/ICASSP49357.2023.10096686. Preprint arXiv:2210.12300 (2022), which is where the key's year comes from; the published version is ICASSP 2023.

[31] Rebecca Schurr Hartley and Roderick A. Suthers. Airflow and pressure during canary song: direct evidence for mini-breaths. *Journal of Comparative Physiology A*, 165(1):15–26, 1989. doi: 10.1007/BF00613795. URL https://doi.org/10.1007/BF00613795. Mini-breaths at 2–27 notes/s, pulsatile expiration at 30–38 notes/s.

[32] Henrike Hultsch and Dietmar Todt. Repertoire sharing and song-post distance in nightingales (*luscinia megarhynchos* b.). *Behavioral Ecology and Sociobiology*, 8(3):183–188, 1981. doi: 10.1007/BF00299828. URL https://doi.org/10.1007/BF00299828.

[33] Henrike Hultsch and Dietmar Todt. Memorization and reproduction of songs in nightingales (*luscinia megarhynchos*): evidence for package formation. *Journal of Comparative Physiology A*, 165(2):197–203, 1989. doi: 10.1007/BF00619194. URL https://doi.org/10.1007/BF00619194.

[34] Dezhe Z. Jin and Alexay A. Kozhevnikov. A compact statistical model of the song syntax in bengalese finch. *PLOS Computational Biology*, 7(3):e1001108, 2011. doi: 10.1371/journal.pcbi.1001108. URL https://doi.org/10.1371/journal.pcbi.1001108. Preprint arXiv:1011.2998 (2010); the repository cites only the preprint.

[35] Stefan Kahl, Connor M. Wood, Maximilian Eibl, and Holger Klinck. BirdNET: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61:101236, 2021. doi: 10.1016/j.ecoinf.2021.101236. URL https://doi.org/10.1016/j.ecoinf.2021.101236. Model weights CC BY-NC-SA 4.0; the BirdNET-Analyzer code is MIT.

[36] Alireza Kazemi, Mariam Kesba, and Pauline Provini. Realistic three-dimensional avian vocal tract model demonstrates how shape affects sound filtering (*passer domesticus*). *Journal of the Royal Society Interface*, 20(198):20220728, 2023. doi: 10.1098/rsif.2022.0728. URL https://doi.org/10.1098/rsif.2022.0728.

[37] Reinhard Kneser and Hermann Ney. Improved backing-off for m-gram language modeling. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 181–184, 1995. doi: 10.1109/ICASSP.1995.479394. URL https://doi.org/10.1109/ICASSP.1995.479394.

[38] Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, 16(12):1289–1296, 2019. doi: 10.1038/s41592-019-0619-0. URL https://doi.org/10.1038/s41592-019-0619-0.

[39] Robert Kubichek. Mel-cepstral distance measure for objective speech quality assessment. In *Proc. IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM)*, volume 1, pages 125–128, 1993. doi: 10.1109/PACRIM.1993.407206. URL https://doi.org/10.1109/PACRIM.1993.407206.

[40] Rodrigo Laje and Gabriel B. Mindlin. Diversity within a birdsong. *Physical Review Letters*, 89 (28):288102, 2002. doi: 10.1103/PhysRevLett.89.288102. URL https://doi.org/10.1103/PhysRevLett.89.288102. A neural-circuit (RA) model driving the syrinx oscillator; not the source of the dimensionless normal form — see amador2008 and sitt2010.

[41] Rodrigo Laje and Gabriel B. Mindlin. Modeling source-source and source-filter acoustic interaction in birdsong. *Physical Review E*, 72(3):036218, 2005. doi: 10.1103/PhysRevE.72.036218. URL https://doi.org/10.1103/PhysRevE.72.036218.

[42] Rodrigo Laje, Timothy J. Gardner, and Gabriel B. Mindlin. Neuromuscular control of vocalizations in birdsong: A model. *Physical Review E*, 65(5):051921, 2002. doi: 10.1103/PhysRevE.65.051921. URL https://doi.org/10.1103/PhysRevE.65.051921.

[43] Ming Li, Xin Chen, Xin Li, Bin Ma, and Paul M. B. Vitányi. The similarity metric. *IEEE Transactions on Information Theory*, 50(12):3250–3264, 2004. doi: 10.1109/TIT.2004.838101. URL https://doi.org/10.1109/TIT.2004.838101. Preprint arXiv:cs/0111054 (2001). Defines the normalised information distance, which is non-computable because it rests on Kolmogorov complexity; cilibrasi2005's NCD is its computable approximation.

[44] Jiali Lu, Sumithra Surendralal, Kristofer E. Bouchard, and Dezhe Z. Jin. Partially observable markov models inferred using statistical tests reveal context-dependent syllable transitions in bengalese finch songs. *Journal of Neuroscience*, 45(9):e0522242024, 2025. doi: 10.1523/JNEUROSCI.0522-24.2024. URL https://doi.org/10.1523/JNEUROSCI.0522-24.2024. The paper argues parsimony and interpretability, not a categorical win over HMMs.

[45] Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction, 2018. URL https://arxiv.org/abs/1802.03426.

[46] Gabriel B. Mindlin and Rodrigo Laje. *The Physics of Birdsong*. Biological and Medical Physics, Biomedical Engineering. Springer, Berlin, Heidelberg, 2005. ISBN 978-3-540-25399-0. doi: 10.1007/3-540-28249-1. URL https://doi.org/10.1007/3-540-28249-1.

[47] Marius Miron, David Robinson, Milad Alizadeh, Ellen Gilsenan-McMahon, Gagan Narula, Emmanuel Chemla, Maddie Cusimano, Felix Effenberger, Masato Hagiwara, Benjamin Hoffman, Sara Keen, Diane Kim, Jane Lawton, Jen-Yu Liu, Aza Raskin, Olivier Pietquin, and Matthieu Geist. AVEX: What matters for animal vocalization encoding. In *Proc. International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2508.11845. Code (MIT) at https://github.com/earthspecies/avex.

[48] Tobias Morocutti, Florian Schmid, Khaled Koutini, and Gerhard Widmer. Device-robust acous-

tic scene classification via impulse response augmentation. In *Proc. 31st European Signal Processing Conference (EUSIPCO)*, 2023. URL https://arxiv.org/abs/2305.07499. Evaluated on acoustic scene classification across recording devices, not on bioacoustics.

[49] Ilyass Moummad, Romain Serizel, Emmanouil Benetos, and Nicolas Farrugia. Domain-invariant representation learning of bird sounds, 2024. URL https://arxiv.org/abs/2409.08589. Preprint.

[50] Brian S. Nelson, Gabriël J. L. Beckers, and Roderick A. Suthers. Vocal tract filtering and sound radiation in a songbird. *Journal of Experimental Biology*, 208(2):297–308, 2005. doi: 10.1242/jeb.01378. URL https://doi.org/10.1242/jeb.01378. The gape regression $A = -5 + 32\log_{10}(B)$ is reported at 6.5 kHz only.

[51] Stephen Nowicki. Vocal tract resonances in oscine bird sound production: evidence from birdsongs in a helium atmosphere. *Nature*, 325(6099):53–55, 1987. doi: 10.1038/325053a0. URL https://doi.org/10.1038/325053a0.

[52] Verena R. Ohms, Peter Ch. Snelderwaard, Carel ten Cate, and Gabriël J. L. Beckers. Vocal tract articulation in zebra finches. *PLOS ONE*, 5(7):e11923, 2010. doi: 10.1371/journal.pone.0011923. URL https://doi.org/10.1371/journal.pone.0011923. Reports gape as a passive filter emphasising frequencies above 5 kHz, not a gape–f0 correlation.

[53] Kazuo Okanoya. The bengalese finch: A window on the behavioral neurobiology of birdsong syntax. *Annals of the New York Academy of Sciences*, 1016:724–735, 2004. doi: 10.1196/annals.1298.026. URL https://doi.org/10.1196/annals.1298.026. Review.

[54] Pritty Patel-Grosz, Salvador Mascarenhas, Emmanuel Chemla, and Philippe Schlenker. Super linguistics: an introduction. *Linguistics and Philosophy*, 46(4):627–692, 2023. doi: 10.1007/s10988-022-09377-8. URL https://doi.org/10.1007/s10988-022-09377-8. A methodological programme; it does not itself assert the song-syntax / call-meaning asymmetry.

[55] Yonatan Sanz Perl, Ezequiel M. Arneodo, Ana Amador, Franz Goller, and Gabriel B. Mindlin. Reconstruction of physiological instructions from zebra finch song. *Physical Review E*, 84(5):051909, 2011. doi: 10.1103/PhysRevE.84.051909. URL https://doi.org/10.1103/PhysRevE.84.051909.

[56] Jeffrey Podos, Joel A. Southall, and Marcos R. Rossi-Santos. Vocal mechanics in darwin's finches: correlation of beak gape and song frequency. *Journal of Experimental Biology*, 207(4):607–619, 2004. doi: 10.1242/jeb.00770. URL https://doi.org/10.1242/jeb.00770.

[57] Simon J. D. Prince and James H. Elder. Probabilistic linear discriminant analysis for inferences about identity. In *Proc. IEEE 11th International Conference on Computer Vision (ICCV)*, pages 1–8, 2007. doi: 10.1109/ICCV.2007.4409052. URL https://doi.org/10.1109/ICCV.2007.4409052. A face-recognition paper. PLDA was adopted by the speaker-verification community later; this is the method's origin, not its speaker-ID use.

[58] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20 (NIPS)*, 2007. URL https://papers.nips.cc/paper_files/paper/2007/hash/013a006f03dbc5392effeb8f18fda755-Abstract.html.

[59] Lukas Rauch, René Heinrich, Ilyass Moummad, Alexis Joly, Bernhard Sick, and Christoph Scholz. Can masked autoencoders also listen to birds? *Transactions on Machine Learning Research*, 2025. URL https://arxiv.org/abs/2504.12880.

[60] Lukas Rauch, Raphael Schwinger, Moritz Wirth, René Heinrich, Denis Huseljic, Marek Herde, Jonas Lange, Stefan Kahl, Bernhard Sick, Sven Tomforde, and Christoph Scholz. BirdSet: A large-scale dataset for audio classification in avian bioacoustics. In *Proc. International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2403.10380.

[61] Tobias Riede and Franz Goller. Peripheral mechanisms for vocal production in birds – differences and similarities to human speech and singing. *Brain and Language*, 115(1):69–80, 2010. doi: 10.1016/j.bandl.2009.11.003. URL https://doi.org/10.1016/j.bandl.2009.11.003.

[62] Tobias Riede and Roderick A. Suthers. Vocal tract motor patterns and resonance during constant frequency song: the white-throated sparrow. *Journal of Comparative Physiology A*, 195(2):183–192, 2009. doi: 10.1007/s00359-008-0397-0. URL https://doi.org/10.1007/s00359-008-0397-0.

[63] Tobias Riede, Roderick A. Suthers, Neville H. Fletcher, and William E. Blevins. Songbirds tune their vocal tract to the fundamental frequency of their song. *Proceedings of the National Academy of Sciences*, 103(14):5543–5548, 2006. doi: 10.1073/pnas.0601262103. URL https://doi.org/10.1073/pnas.0601262103.

[64] David Robinson, Marius Miron, Masato Hagiwara, Benno Weck, Sara Keen, Milad Alizadeh, Gagan Narula, Matthieu Geist, and Olivier Pietquin. NatureLM-audio: an audio-language foundation model for bioacoustics. In *Proc. International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2411.07186.

[65] Christian Rutz, Michael Bronstein, Aza Raskin, Sonja C. Vernes, Katherine Zacarian, and Damián E. Blasi. Using machine learning to decode animal communication. *Science*, 381(6654):152–155, 2023. doi: 10.1126/science.adg7314. URL https://doi.org/10.1126/science.adg7314. Perspective.

[66] Tim Sainburg, Marvin Thielk, and Timothy Q. Gentner. Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLOS Computational Biology*, 16(10):e1008228, 2020. doi: 10.1371/journal.pcbi.1008228. URL https://doi.org/10.1371/journal.pcbi.1008228.

[67] Pratyusha Sharma, Shane Gero, Roger Payne, David F. Gruber, Daniela Rus, Antonio Torralba, and Jacob Andreas. Contextual and combinatorial structure in sperm whale vocalisations. *Nature Communications*, 15:3617, 2024. doi: 10.1038/s41467-024-47221-8. URL https://doi.org/10.1038/s41467-024-47221-8. At least 143 frequently realised coda combinations; the figure "156" does not appear in the paper.

[68] Jacobo D. Sitt, Ana Amador, Franz Goller, and Gabriel B. Mindlin. Dynamical origin of spectrally rich vocalizations in birdsong. *Physical Review E*, 78(1):011905, 2008. doi: 10.1103/PhysRevE.78.011905. URL https://doi.org/10.1103/PhysRevE.78.011905.

[69] Jacobo D. Sitt, Ezequiel M. Arneodo, Franz Goller, and Gabriel B. Mindlin. Physiologically driven avian vocal synthesizer. *Physical Review E*, 81(3):031927, 2010. doi: 10.1103/PhysRevE.81.031927. URL https://doi.org/10.1103/PhysRevE.81.031927.

[70] Roderick A. Suthers. Contributions to birdsong from the left and right sides of the intact syrinx. *Nature*, 347(6292):473–477, 1990. doi: 10.1038/347473a0. URL https://doi.org/10.1038/347473a0.

[71] Toshitaka N. Suzuki, David Wheatcroft, and Michael Griesser. Experimental evidence for compositional syntax in bird calls. *Nature Communications*, 7:10986, 2016. doi: 10.1038/ncomms10986. URL https://doi.org/10.1038/ncomms10986. Reversed order (D-ABC) produced a reduced response, not no response.

[72] Evren C. Tumer and Michael S. Brainard. Performance variability enables adaptive plasticity of 'crystallized' adult birdsong. *Nature*, 450(7173):1240–1244, 2007. doi: 10.1038/nature06390. URL https://doi.org/10.1038/nature06390. The "∼1 % f0 CV" figure the repository attributes to this paper could not be located in accessible text.

[73] Bart van Merriënboer, Vincent Dumoulin, Jenny Hamer, Lauren Harrell, Andrea Burns, and Tom Denton. Perch 2.0: The bittern lesson for bioacoustics, 2025. URL https://arxiv.org/abs/2508.04665. 1536-d embedding, EfficientNet-B3 trunk, 14 795 output classes of which 14 597 are species, 5 s at 32 kHz; code Apache 2.0.

[74] William E. Wood, Peter J. Osseward, Thomas K. Roseberry, and David J. Perkel. A daily oscillation in the fundamental frequency and amplitude of harmonic syllables of zebra finch song. *PLOS ONE*, 8(12):e82327, 2013. doi: 10.1371/journal.pone.0082327. URL https://doi.org/10.1371/journal.pone.0082327. True source of the $R^2 \approx 0.03$ figure the repository credits to a non-existent "Chen et al. 2013".

[75] Xeno-canto Foundation. xeno-canto: sharing wildlife sounds from around the world. https://xeno-canto.org. Per-recording Creative Commons licences; accessed 2026-08-01.

[76] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with MixStyle. In *Proc. International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2104.02008. Extended journal version: Zhou et al., *MixStyle Neural Networks for Domain Generalization and Adaptation*, IJCV 132(3):822–836, 2024, https://doi.org/10.1007/s11263-023-01913-8.

[77] Sue Anne Zollinger, Tobias Riede, and Roderick A. Suthers. Two-voice complexity from a single side of the syrinx in northern mockingbird *mimus polyglottos* vocalizations. *Journal of Experimental Biology*, 211(12):1978–1991, 2008. doi: 10.1242/jeb.014092. URL https://doi.org/10.1242/jeb.014092.

A The complete pre-registered gate ledger

Table 12 lists every gate the project registered before its run, the outcome, and the verdict. It is complete: refuted arms are not omitted, and arms whose machinery was subsequently deleted keep their row, because the measurement is what survives a deleted flag. Of 58 gates, 20 passed, 37 ended in a written refusal, and 1 is unrun. That count is reproduced from this table by `analysis/ledger_tally.py`, which states its classification rule before it counts: a verdict naming an adoption is a pass unless it also says the gate failed, so "fails; ships as an honesty gate" is a refusal. Four refusals still produced an asset: the map-flip lever became an honesty gate, the mel target became the verdict instrument for future fit gates, label assignment became a build flag, and the source-axis deflation produced the finding that the recorded/synthesised difference is per-type geometry, not one direction.

Where a row is marked *debug* the measurement is on the 100-species debug subset, the 37-species store, or a similarly sized sample; those numbers license arm-against-arm comparison and not absolute claims.

Table 12: The complete ledger of pre-registered gates.

what was tried	the gate, set in advance	what it measured	verdict
Typing in the critic's 32-d embedding	placement up, cosine guardrail held	38.4 % → 51.7 %, cosine 0.807 → 0.796 *(debug)*	**ships**
Parallel and streaming learn	bit-identical to serial	bit-identical, ∼3 h → ∼15 min	**ships**
Fricative route: noise through fitted formants	resynthesis beats the tonal fit, per type	bench 0.584 → 0.656, 13 wins / 0 losses	**ships per type**
Tracheal comb per species	comb improves the resynthesis	won 10 of 22 species	**ships where it wins**
Measured texture on authored one-tone calls	gain against the same bird's learned types	14 of 15, median +0.148	**ships**
Respiratory oscillator, hand-set default	beat the envelope on median cosine	0.343 → 0.336	fails, parked
Respiratory oscillator, per-species fit	win the type, gate held everywhere	won 445 of 12 680 scored types on the cosine. **Re-judged on the bench and lost**: 33.1 % of 175 types improved, median −4.18 dB MCD, despite +0.034 median cosine	**ships**, and the gate that shipped it is now known to read fidelity backwards
Two-voice coupling, hand-set default	beat the single voice	0.343 → 0.294	fails, parked
Two-voice coupling, per-type fit	win the type	won 2 027 of 5 922 screened. **Re-judged on the bench and held**: 55.6 % of 1 737 types improved, median +1.13 dB MCD, sign test $p = 3.2 \times 10^{-6}$	**ships per type**
Fixed-waveguide tract	placement and cosine both up	+6.2 pts on 5 species; regression over 22 (23.1 → 22.2 %)	refuted, gated off
Measured partial series as fit evidence	cosine and placement up	cosine 0.776 → 0.768, placement 61.7 → 55.6 %	refuted
Global spectral-tilt correction	placement up	tilt corrected exactly, placement 23.1 → 20.2 %	refuted
Realism logit as the fit objective	placement and cosine hold	cosine 0.762 → 0.335, placement 17.3 %; replicated at 40 k	**refuted twice**
Mel-path objective at $w = 0.5$	≥ 60 % MCD wins, ΔMCD > +1 dB, cosine drop ≤ 0.02 / max 0.05	31 % wins ($n = 80$), 0.0 dB, max drop 0.120	fails and not adopted
Mel-path as a *measurement*	$\lvert r\rvert \geq 0.4$ against MCD	Spearman −0.456, Pearson −0.346 ($n = 1\,200$; was −0.420/ − 0.305 at $n = 200$)	**adopted as the bench**
Path-cosine mel form	see the damaged-half fixture	drop 0.017 against exp(−LSD)'s 0.540	refuted
RFF ridge browser bridge	median held-out cosine ≥ 0.8	0.962 on the debug build	**ships**
Linear ridge bridge, the incumbent	the same 0.8 gate	0.7997 after the embedding grew finer	refuted, replaced
Source-split cluster seeds	type coverage up, purity no worse	coverage 71.0 → 93.1 % on the dumped matrix, 94.6 % in the build; purity 48.1 → 56.6 %	**ships**
Free extra cluster seeds	the same	coverage 71.0 → 78.8 % (+144 seeds) and 86.9 % (2× types)	superseded by the above
Chunk mining evidence gate	PMI $\cdot\sqrt{\text{count}} \geq 6$, shuffled control finds nothing	control finds 0 chunks; aliasing faked PMI to 1.2 without the gate	**ships**
Corpus-breadth ingest vs the four placement gates	species > 35 %, type ≥ 56 %, cosine ≥ 0.74, $\lvert f_0\rvert \leq 5$ %	species 24.6 → 11.5 % (chance 2.8 → 0.9 %); other three held	**fails** on species
Channel-EQ + Freq-MixStyle augmentation	−30 % recordist lift, species held	−0.7 % recordist *(debug)*	fails

continued overleaf

Table 12, continued

what was tried	the gate, set in advance	what it measured	verdict
Recordist-adversarial head	the same	−33 % recordist for −14.7 pts species accuracy *(debug)*	fails
Within-species adversarial head	the same, species-safe by construction	species flat to +10 %, recordist −6 % *(debug)*	fails
Naive two-domain merge	cross-domain retrieval ≥ +10 pts	+14.4 pts at 12 species, +3.6 at 63	no adoption; direction holds
Cross-domain contrastive positives	≥ +10 pts over the merge	+13.4 at 12 species, +1.7 at 63	fails on re-test
Centre / whiten by recordist	recordist lift halved, species not falling	recordist 0.1665 → 0.1687 and 0.1540; species −32 to −42 %	refuted
Harmony by recordist	the same	recordist −71 %, species −70 %	fails, 1:1 trade
Correction by recording	the same	both removed (−74 / −75 %); labels nearly nested	fails, uninformative
T1: external space vs the recordist confound (Perch)	recordist lift −30 % relative	0.0891 → 0.0744, i.e. −16.5 %	**fails**
T1 with BirdNET	the same	0.0891 → 0.1107, i.e. +24 %	**fails harder**
T3: external species retrieval (Perch)	beat the critic's species lift	0.0074 → 0.0588 (8×)	**passes**
T3 with BirdNET	the same	0.0074 → 0.0393 (5.3×)	passes
T2: external novelty AUC	beat 0.846 / TPR 32.2 %	not run	unrun
Jin–Kozhevnikov repeat-count grammar	beat first-order on held-out bouts	−0.348 nats/token, 0 of 30 species *(debug)*	refuted
Hidden-state HMM grammar	the same	+0.422 nats overall, all of it in the 10 % unseen tail *(debug)*	refuted on decomposition
Kneser–Ney unseen-transition backoff	gain on unseen contexts, no loss on seen	+1.90 / +2.32 nats, seen side exactly 0.0000	**ships**
HMM re-run after the depth harvest	≥ +0.1 nats over the repaired first-order	≈ +0.37 nats, 78 of 104 species	wins the rule; **not adopted**
Song-level edit-distance kernel	beat bag-of-types on species lift, recordist lift < 0.05	+0.877 vs +0.924; recordist +0.264	fails both gates
Menzerath–Altmann law on the inventory	separate the corpus from its own shuffle	median exponent +0.019 where the law needs negative	the law does not hold
Normalised compression distance	the same	control separates species better (0.038 vs 0.018)	refuted
Second-order structure, plug-in estimator	k=1 → 2 drop above control's	invalid comparison; curves start 2.4 bits apart	retracted by the project
Second-order structure, held-out CE	median ≥ +0.05 bits/token, ≥ +0.05 over shuffle, share ≥ control +20 pts	−0.091 bits/token; share 25.3 % vs 4.1 %	**fails** (magnitude); share passes
Song-type detection vs a permutation null	recurrence above the null, control ≤ 0.2 %	control 0.170 % at the shipped constants	**ships**
Repeat classes	≥ 90 species with a song, control ≤ 0.2 %	98 species, 356 types, control 0.170 %	**adopted**
Per-species recurrence bar	control returns to 0.2 %	no candidate threshold reaches it (0.266–0.442 %)	refuted, removed
Per-species bout rule	the same	105 species at 0.628 % control yield	refuted, removed
Session gate: off-manifold 5-NN detection	AUC ≥ 0.90	0.953, and ≥ 0.92 from 256 shipped anchors	**ships**
Session gate: map-flip lever	+5 points of agreement	+3.5 points raw; tie-adjusted +8.7 to +13.7	fails; **ships as an honesty gate**
Session gate: prune bad sessions from the fit	evaluation cosine cost ≤ 0.005	−0.0147 and −0.0062 on two stores	refuted
Fewer rendered member variants	seeds per type toward 1.0, purity and coverage no worse	2.000 → 1.993; purity 0.423 → 0.568	fails on the stated gate
Source-axis deflation	the same	42 % of the space's energy removed, source gap 7.233 → 0.0, seeds/type 1.999	refuted
No synthetic point for a type with members	the same	seeds/type exactly 1.000, k 4 298 → 2 155; retrieval +0.02	**adopted, on size**
Label assignment vs uncapped k-means	type retrieval ≥ k-means +10 pts	+2.5 pts uncapped; +14.2 at production's shed ratio	**fails as written**; ships as a flag
Learned inverse, first bake-off	2× faster at equal cosine	0.737 vs 0.734, and 0.2 s either way	fails its ship rule
Learned inverse as an initialiser	median +0.02, ≥ 60 % of types improved, no species worse than −0.05	+0.0135, 52.8 %, worst species −0.0816	**fails on all three**
Two-start descent	median +0.02 with *no* species worse at all	+0.0330; 0 of 1 661 types and 0 of 100 species worse; cost 61 min vs 11 *(debug)*	**passes**
first attempt, selection before the fricative refinement	the same	+0.0320 but 48 types worse (to −0.049), 100 % of them fricative-flagged	failed, bug found

B Corpus, stores and the built asset set

The current production build holds 348 461 points in 14 003 regions, 13 903 type clusters with zero source splits, plus 100 k-means clusters for the 20 241 points that carry no label. At 100 % type coverage, built in 1 136.8 s with every verification check passing. Its three-dimensional projection reproduces high-dimensional distances at Pearson 0.691 and Spearman 0.689 over 200 000 sampled pairs. Centroid assets are 7.69 MB of JSON plus a 3.81 MB binary sidecar, against the 44 MB whole-file parse that blocked an earlier build. Cluster purity reads 94.5 % and is *not* a measurement on this build. It is 1.0 by construction on the labelled clusters, and the manifest says so.

A separate scaling result belongs here because it is a reproducible engineering fact rather than a finding about birds. The cross-species relations document is quadratic in species count: at 2 421 species it is 2 929 410 unordered pairs and 760 MB, of which only 12 762 pairs (0.44 %) carry a shared type at all. The shipped form writes each species' eight nearest partners, deduplicated, 14 425 pairs, 5.6 MB, a 135-fold reduction. It was checked against the complete document across all 2 421 species at the application's query limit with zero mismatches. The file declares the retention rule and both totals, and states in as many words that the kept shared-type count is not the corpus's shared-type count.

A data-integrity finding about that same document, because it reached the interface before it was caught. The nearest species pairs are the part of a similarity asset a reader trusts most, being the part shown first. In the shipped 14 488-pair document, *four of the five nearest pairs were the same recording under two species labels.* The clearest is a pair at a distance of −0.0: two iNaturalist observation identifiers resolving to byte-identical audio, 965 770 bytes with the same SHA-256 and the same rights holder, each carrying one recording and 146 syllables, every row identical once the identifiers are stripped, all eight of eight type centroids bit-identical. Its "eight shared types" is one repertoire counted twice, and its distance is float overshoot on a unit vector against itself rather than a clamp.

It is a class of row and not one row: 17 cross-species duplicate groups over 35 recordings and 40 species in the harvest store. The cause is upstream and simple — the multi-source ingest de-duplicates on the source's recording identifier and no content hash exists anywhere in the harvester, so the same bytes under two observation identifiers pass straight through. The downstream guard applied here refuses a pair when more than half of one species' sampled syllables have a bit-identical twin in the other, symmetrically, which costs 13 pairs and 27 % of the shared-type candidates and removes all four artefacts from the head of the list; the new nearest pair is a genuine one at 0.0311. The count and the rule are written into the asset, because an exclusion nobody can see reads as a corpus with no duplicates in it. The durable fix, a content hash at ingest, is not yet applied and would refuse those 35 recordings on the next harvest.

Why it is worth a paragraph in a methods paper. Nothing in the pipeline was wrong: the distance was computed correctly, the threshold was derived correctly, and the assertion that no species has a within-species nearest-type distance below 10^{-4} still holds. The failure was that a corpus assembled from overlapping open archives contains the same audio twice, and a similarity measure over it will find that first and rank it highest. Any comparative claim drawn from archive-scale bioacoustic data has this exposure, and a distance of exactly zero between two species is the symptom to look for.

C The oscillator and its fitted parameters

The source is the normal-form flapping oscillator, integrated by forward Euler at 16 to 64 times the audio rate:

$$\dot{x} = y, \tag{1}$$

$$\dot{y} = -\alpha\gamma^2 - \beta\gamma^2 x - \gamma^2 x^3 - \gamma x^2 y + \gamma^2 x^2 - \gamma x y, \tag{2}$$

with x the labial position, y the labial velocity, α the air-sac pressure, β the labial tension and γ a pure time-scale. This is the published normal form term for term, with a sign convention on x that places the limit cycle at $\alpha, \beta > 0$ where the literature states the onset threshold as $\alpha < 0$; the equivalence was verified numerically. Both implementations start from $(x, y) = (10^{-4}, 0)$ and bounds-check the state once per output sample instead of clamping every integration step, because a tight per-step clamp flattens the peaks of the large-amplitude regimes the model legitimately reaches at high tension.

Because rescaling time by γ leaves the equations unchanged, γ is a pure time-scale and

$$f_0 = \gamma\,\Phi(\alpha, \beta), \qquad \Phi(0.1, 1) = 0.17061 \text{ (measured)}, \tag{3}$$

so the calibration table is measured once at $\gamma = 24\,000$ and read at any other γ by scaling the target frequency. Per syllable, $\gamma = \mathrm{clamp}(f_{0,\mathrm{ref}}/0.17061,\ 1200,\ 150\,000)$ with $f_{0,\mathrm{ref}}$ the geometric mean of the syllable's contour frequencies.

The tract is a two-pole resonance centred on the instantaneous f_0, plus up to four fixed resonances at absolute frequencies from the tracheal comb

$$\begin{aligned} f_n &= (2n-1) f_1, \quad n = 1 \ldots 4, \\ \text{gain} &\propto 1/(2n-1), \quad Q = 6, \end{aligned} \tag{4}$$

a stopped quarter-wave series with f_1 measured once per species. Turbulence is added to the source before the tract, scaled by airflow. Roughness multiplies *tension*,

$$\beta \leftarrow \beta \cdot \mathrm{clamp}(1 + 0.12\,\rho\, w\, G,\ 0.5,\ 1.5)\,, \tag{5}$$

Table 13: The measurement stores. Each row is a fixed corpus, and arms are comparable only within a row. The last two rows are the current production state.

store or phase	syllables	species	what it added
original harvest	39 976	22	the reference corpus for the naturalness work
first patch-carrying store	9 041	25	the learned-typing bake-off store
the same at 40 k	39 993	25, 13 after a fixed trim bug	the +13.3-point replication
+ multi-source ingest	39 874	121	the first non-xeno-canto fetch
+ iNatSounds	226 264	3 811	6 453 recordings in one pass, zero failures
+ **depth harvest**	**340 079**	**3 811**	+139 440 syllables, 7 992 → 10 184 recordings
+ under-floor top-up	378 224	3 812	1 500 recordings, segmentation band applied
current fit	**367 196 typed**	**2 650 fitted, 16 283 types**	median resynthesis cosine 0.699

with w a xorshift32 noise sequence one-pole lowpassed to 140 Hz and G normalising it to unit standard deviation. Acting on tension, not on pressure or on the output is the one point the literature is specific about [2], and it makes the frequency and amplitude wobble arrive correlated through the oscillator's own dynamics, which is what separates a rough voice from a clean voice with hiss added.

The optional respiratory path replaces the designed envelope with the published two-state air-sac dynamics [18],

$$\tau_x \dot{x}_s = -(1 + x_s^2)x_s - p + F_0 f(t), \tag{6}$$

$$\tau_p \dot{p} = -(1 + x_s^2)x_s - \big(1 + \alpha_r(p)\big)p + F_0 f(t), \tag{7}$$

with x_s the air-sac wall displacement, p the gauge pressure, α_r a Heaviside conductance giving inflow and outflow different damping, and the tract driven by $\sqrt{p_s}\, y$. The inertive low-frequency limit of Laje and Mindlin [41], not the acceleration injection some reference implementations use. It is off by default.

Thirteen gesture axes are searched per type by coordinate descent; the measured pitch contour and duration are excluded from the search, because the pitch tracker reaches them to about 1 % and a search must not overwrite a measurement.