Calibrating exchange-level claims about sperm whale codas against surrogates: overlap, ornamentation and rubato under independent pre-registered control

Andrew Cordivari · Independent Researcher · andrew.cordivari@gmail.com

Preprint, 2026-09-08. Not peer reviewed.

Code and data: `github.com/acordivari/ling`, tag `preprint-v1` (MIT, public). Every input is fetched from the original authors' own deposits and none is redistributed (§8); every statistic is deterministic and reproduces byte-for-byte.

Abstract

Sharma et al. (*Nat. Commun.* 15:3617, 2024) report that sperm whale codas carry a combinatorial "phonetic alphabet" built from four dimensions, two of which — rubato (smooth modulation of coda duration across an exchange) and ornamentation (an extra click marking sequence edges) — are properties of coda *sequences* rather than of codas. Sequence-level claims presuppose a model of the exchange itself, and the prevailing model — conversation-like alternation — has recently acquired published support in the form of next-coda predictability [3, 4]. Twenty-seven months after publication we find no independent statistical evaluation of these claims in the indexed literature, and the PMC record carries no Matters Arising, Comment or correction; we state that as a search result rather than as a fact about what exists. We pre-registered and ran three tests on the authors' own public deposit. First, the exchange itself. Sharma et al. describe chorusing explicitly — their glossary defines an exchange as a period in which more than one whale calls, and they state that interacting whales produce codas "both alternately (i.e., turn-taking) or near-simultaneously (i.e., overlapping)." What has not been done is to calibrate that description against surrogates. We find that **overlap is far more common than any envelope- or bout-preserving surrogate produces**: on a label-free statistic that no relabelling can move — all overlapping adjacent pairs over all adjacent pairs — the corpus gives 25.4% against 15.6% under speaker rotation and 16.1% under ±2 s jitter (permutation floor p = 0.0005). **Turn-taking itself, scored on the authors' own definition, is far below chance.** Sharma et al.'s glossary defines it as codas beginning "within two seconds, but after the termination" of the previous one, and requires the pair to be cross-speaker. Scored over all adjacent pairs — a denominator that moves under 0.5% across surrogates — the rate is **0.1360** (153/1,125) against **0.2337** under a permutation of each whale's own inter-coda intervals, the best-behaved null available (z = −6.97, p = 0.0005); rotation and jitter give larger z values but inflate same-speaker within-window mass by 40% and 238% respectively and are read as upper bounds. Eleven of twelve deployments fall below their own null (sign test p = 0.0063). Two calibration controls support the reading: on a corpus built to have zero cross-whale contingency the interval-permutation null returns z = −0.19 ± 1.01 while rotation and jitter return −4.43 and −2.97, so only the former is unbiased; and injecting genuine responders moves z monotonically from −7.04 to +12.75, so the statistic detects them when they are there. This is not the overlap result restated: adding overlapping pairs back in leaves the combined near-simultaneous mass *also* below chance (0.4071 against 0.4395 under the same null, z = −3.99). Both results in this section assume the animals were within roughly 750 m on average: the deposit records arrival times, and a constant propagation delay of ±1 s would carry both to their nulls. Whales on this corpus either overlap or wait appreciably longer than independence predicts, and the intermediate band is depleted at both ends. Two statistics we registered — a switch rate and a median switch gap over admissible pairs — turn out not to

be scoreable, because every surrogate that moves timestamps changes how many pairs are admissible, and admissibility is the complement of overlap; we report that as a design limitation rather than as a null. Second and third, the two sequence-level alphabet dimensions, using the authors' own per-coda rhythm classifications and ornamentation flags (alignment validated at 95.65% with all mismatches confined to one residual class), under two controls the original analyses do not apply jointly: stratification by rhythm-class composition, and a sweep of the sequence-segmentation threshold. Both tests were pre-registered with the prediction that neither feature survives; the prediction was half wrong. Ornamentation's positional association does not meet the pre-registered criterion and is sensitive to segmentation: it appears only at cuts of 10–15 s — bracketing the 10 s window inside the flag's own operational definition — and rhythm-class composition accounts for a third to half of it where it appears. Rubato holds throughout: within-class duration drift exceeds composition-preserving nulls at every segmentation cut ($z = 6.4$–19.4, permutation floor $p = 0.0005$), is robust to leave-one-recording-out, and holds inside single rhythm classes, where click count is invariant and duration is pure inter-click timing. Smooth tempo modulation in sperm whale codas is real structure. A registered follow-up (experiment 10) then tested its provenance: tempo deviations of nearby codas from different whales covary ($r = 0.30$) — a cross-whale signature of the kind the paper's "perceived and controlled" reading rests on — but an envelope-preserving jitter null with zero cross-whale coupling reproduces the covariation entirely ($p = 0.48$–0.97 at every window). The published evidence for communicative use of rubato is not re-established once these two additional controls are applied, at the reported discrimination; within-whale persistence is scale-limited and concentrated in a single annotation scene. Rubato stands as structure; its use remains undemonstrated.

---

1. Introduction

Sharma et al. [1] analyse 8,719 codas from the Dominica Sperm Whale Project and propose that coda structure decomposes into rhythm, tempo, rubato and ornamentation. The first two are per-coda features with a long prior literature [2]. The novel claims are the two *context-sensitive* dimensions:

- **Rubato:** duration varies smoothly across consecutive codas — adjacent same-type codas from the same whale differ less in duration than chance (0.05 s vs 0.08 s, $p = 0.0001$, $n = 2{,}593$), drift direction is sustained across sequences, and overlapping codas from different whales match in duration (0.099 s vs 0.129 s, $n = 908$). The authors evaluate "whether rubato is perceived and controlled by measuring whales' ability to match their interlocutors' coda durations when chorusing."
- **Ornamentation:** ~4% of exchange codas carry an extra click, occurring disproportionately at sequence beginnings (Fisher's exact, OR 2.00, $p = 0.0006$) and ends (OR 1.71, $p = 0.008$).

These claims matter beyond cetology: they are the empirical basis for the "phonetic alphabet" framing that has organised subsequent machine-learning work on whale vocalisation. As of August 2026 the PMC record shows no Matters Arising, Comment or correction. We note that the journal's Matters Arising window — 18 months from publication — closed in about November 2025, so the absence of a formal comment is partly a function of that deadline rather than of consensus. This paper is therefore offered as a standalone re-analysis rather than as a comment on the original.

Both sequence-level dimensions rest on a model of the exchange. Sharma et al. are explicit that they do not assume alternation: their glossary defines an exchange or chorus as "period of time where codas are made by more than a single whale," and their results state that "across a chorus, interacting whales produce codas both alternately (i.e., turn-taking) or near-simultaneously (i.e., overlapping)." Their rubato analysis is built on 908 overlapping cross-whale pairs. What is missing is not the observation but its

calibration: no published analysis of this corpus asks how much overlap a surrogate with no cross-whale coupling would produce. That gap matters because the conversational reading has since been amplified downstream, where it has acquired support of a specific kind — *predictability*. Assayag et al. [3] report that the structure of the next coda from one whale in a Dominica pair is predictable from up to ten preceding codas of the other, using a "Coda Break" metric signed so that negative values quantify overlap. A sequence model trained on the same annotation corpus analysed here [4] reports order dependence and long-range dependencies on up to eight preceding codas; its preprint abstract described the exchanges as exhibiting "predictable turn-taking," a phrase the camera-ready abstract no longer carries — there, turn-taking enters only as a category within the model's input tokens, not as a tested claim about exchange timing. Predictability, however, is compatible with two very different generating processes: alternating response, and simultaneous calling driven by a shared activity envelope. No published analysis of this corpus separates them with surrogates. We therefore test the exchange model first, before testing the two dimensions that presuppose it.

Two features of the corpus make an independent control both necessary and unusually clean. Necessary, because the two context-sensitive claims are *sequence-level* statistics, and sequence-level statistics inherit two well-known failure modes: composition (which coda types occur where) and segmentation (how continuous click trains are cut into "sequences"). The paper's glossary fixes that parameter at **8 s** — "a sequence of calls made by a given whale where every consecutive pair of calls occur within 8 seconds" — a value the machine-readable deposit does not carry, and which is distinct from the 10 s window used to flag ornaments. Clean, because the authors published their own per-coda classifications: rhythm-class indices and ornamentation flags aligned 1:1 with the public dialogue corpus. We therefore test the authors' own labels, not a reimplementation of their definitions — removing the most common way an independent check goes wrong.

We pre-registered both tests, including the prediction that **neither** feature would survive, before computing any test statistic. The prediction was wrong about rubato. We consider that asymmetry — one claim not holding under these controls while the other strengthens — the most informative outcome this design could produce, and more credible than either a uniform negative or a uniform confirmation.

2. Data, alignment, and three traps

Corpus. `sperm-whale-dialogues.csv` from the authors' deposit: 3,840 codas with recording id, click count, duration, per-click ICIs, a per-recording speaker index, and absolute onset time. (The paper describes Dataset 2 as 3,948 codas; the public deposit carries 3,840. We flag the 108-coda bookkeeping difference without interpreting it.) Labels: `ornaments.p` (binary flag per coda) and `rhythms.p` (18-class index per coda), both length 3,840, plus the 18 class centroids.

Alignment gate (G1). The claim "`rhythms.p[i]` describes CSV row *i*" has an independent test: each centroid has a click count, which must match the row's own `nClicks`. Agreement is 3,673/3,840 = 95.65% — and all 167 mismatches sit in class 17. A detection rule (a class whose members' click counts mostly disagree with its own centroid is a bucket, not a category) finds class 17 and nothing else: it holds 175 codas with 1–29 clicks under a 10-click centroid, while class 14 — same centroid — holds 13 codas, all exactly 10 clicks. Class 17 is a residual bucket and is excluded; it holds 13 ornamented codas that would otherwise let unclassifiable material carry positional signal. A consequence used throughout: **within every non-residual class, click count equals the centroid's exactly**, so within-class duration variation is inter-click timing, never click count.

Zero-duration trap (G2). Eight codas have duration exactly 0; all are single clicks (no inter-click interval, hence no rhythm and no duration sequence), and all sit in class 17.

Dual-tag trap (G0). The `a`/`b`/`c` suffix in recording ids is a DTag letter, not a session: deployments carry several tags at once, so the same acoustic scene can appear twice with independently assigned speaker indices (constant clock offsets between duplicated scenes: 1,449.7 s and 1,870.2 s). We keep one tag per deployment (the one contributing the most codas): 3,083 of 3,840 codas retained. The final observation universe is 2,958 classifiable codas, 116 of them ornamented.

3. The composition confound, measured before testing

Ornamentation is strongly associated with rhythm class. In the full corpus, two classes carry most of the flag: class 10 (rate 0.667) and class 16 (0.596), against 0.006 in the majority class (n = 2,359). Rhythm class alone predicts the ornament flag with a 19.2% error reduction over the majority-class baseline. Sharper still: **140 of 151 ornamented codas have exactly their class's centroid click count** — the "extra" click is absorbed by the class assignment, so classes 16 and 10 behave as ornamented variants in their own right.

The design consequence is immediate: a positional test of ornamentation that does not stratify by rhythm class is a test of where classes 16 and 10 occur. The confound is strong but not total (their rates are 0.6–0.67, not 1.0), so a stratified test has leverage rather than being vacuous — we report that leverage (259.7–429.8 by the harmonic-mass convention, against the floor of 33.3 this project carries) beside every p-value.

4. Methods

S0 (exchange structure). Run first, because S1 and S2 presuppose its answer. The design turns on a structural fact: a whale cannot overlap itself, so every pair of codas that overlap in time is necessarily attributed to two animals, and before any filtering 40.55% (908/2,239) of adjacent cross-speaker pairs in this corpus overlap. Those are switches forced by simultaneity, carrying no information about turn-taking. We therefore score two statistics over *admissible* pairs only — a pair is admissible when coda i+1 begins after coda i ends, so that both a same-speaker and a cross-speaker continuation were physically possible: (1) the **switch rate**, the fraction of admissible adjacent pairs where the speaker changes; and (2) the **median switch gap**, onset-to-onset time at admissible switches. Overlap itself is scored *before* the admissibility filter — it is precisely the quantity the filter removes. Three surrogate families, each controlling what the previous ignores: **N1**, naive label shuffle (destroys everything; included to show what it inflates); **N2**, speaker rotation — a rigid random shift of one speaker's entire timeline, preserving every within-speaker interval (the primary null). Rotation does **not** eliminate self-overlap: at the wrap seam, and under every other family, the surrogates manufacture anatomically impossible self-overlaps at 10.1 to 141.7 per draw against zero in the corpus (§5.1), which a statistic scored over cross-speaker pairs discards silently; **N3**, local jitter, swept over windows of ±2, 5, 10, 20, 45 and 90 s — an effect that is genuine response contingency must survive small windows, where the surrogate still shares the observed co-activity envelope; an effect that appears only at large windows is co-activity, not contingency. Pre-registered prediction: the switch rate is null under N2; no strong prediction for the gap. What would falsify it: a switch rate surviving N2 and N3 at small windows.

Scope of the exchange universe. S0 is scored only on recordings carrying exactly two speaker indices and at least 6 codas — 65 of 219 recordings, 1,192 of 3,083 post-G0 codas. Exchanges among three or

more whales are outside this analysis entirely, and nothing here generalises to them. Sharma et al.'s chorus definition ("more than a single whale") is broader than the dyads tested here.

Sequences. The machine-readable deposit carries no sequence definition, though the paper's glossary fixes one at 8 s (§1). We define a sequence as consecutive codas by the same speaker within one recording, separated by silence (offset-to-onset) of at most GAP seconds, and sweep GAP over {3, 5, 10, 15, 30} s as a pre-registered falsification device: a statistic that a segmentation parameter can switch on and off is measuring the cut. Same-speaker sequences are deliberate, and S0's result (§5.1) is the licence: overlap far exceeds chance, so a cross-speaker "exchange" pools two animals whose calls frequently coincide — and mixing two animals with different duration means would manufacture sequence-level duration structure outright.

S1 (ornamentation). Δ = P(ornamented | sequence-final) − P(ornamented | non-final), two arms with the identical observed statistic: a naive null (permute flags anywhere) and a stratified null (permute flags within rhythm class — destroying position, preserving composition and per-class rates). Two-sided.

S2/S2b (rubato). Pooled per-sequence-centered lag-1 autocorrelation of duration. S2's null shuffles durations within sequence (destroying order, preserving each sequence's duration multiset); S2b's shuffles within sequence × class (additionally preserving the class ordering and any duration structure it carries). One-sided: smoothness asserts positive autocorrelation. Excluded codas break sequences into fragments rather than being spliced over; fragments require ≥ 3 codas.

Discipline. Both statistics, their nulls, the sweep, the survival criterion ($p < 0.05$ at *every* GAP with mass above floor), and the prediction that neither survives were fixed before any statistic was computed. Negative controls (placebo draws from each arm's own null pushed through the full pipeline, 40 repeats per arm) fired at 1–5/40 against a nominal 2/40. Everything is seeded and byte-for-byte reproducible.

5. Results

5.1 Overlap far above chance; alternation and latency not adjudicated

Universe after G0: 65 two-speaker recordings, 1,192 codas, 839 admissible pairs, 448 admissible switches (3,083 of 3,840 codas retained by the gate).

null	overlap	z	p	switch rate	z	p
(observed)	**0.390**			**0.534**		
N1 naive label shuffle	0.256	10.4	**0.0005**	0.499	2.2	0.0190
N2 speaker rotation	0.212	10.8	**0.0005**	0.646	−8.1	1.0000
N3 jitter ±2 s	0.252	9.6	**0.0005**	0.538	−0.3	0.6212
N3 jitter ±5 s	0.235	11.1	**0.0005**	0.533	0.0	0.4958
N3 jitter ±10 s	0.231	11.1	**0.0005**	0.510	1.7	0.0370
N3 jitter ±20 s	0.219	12.0	**0.0005**	0.491	3.0	0.0025
N3 jitter ±45 s	0.199	12.7	**0.0005**	0.428	7.1	0.0005
N3 jitter ±90 s	0.184	12.4	**0.0005**	0.335	13.0	0.0005

Two properties of this ladder have to be stated before any of it is read (`tools/exp08_corrections.mjs`). First, **every surrogate manufactures anatomically impossible self-overlaps** — a whale overlapping itself — at 10.1 to 141.7 per draw against zero in the corpus. The rate

above is scored over cross-speaker pairs, so a surrogate that invents self-overlaps disposes of them silently. Second, **N1 cannot move a label-free overlap statistic at all**: shuffling labels does not move timestamps. We therefore also report the anatomy-neutral rate — all overlapping adjacent pairs over all adjacent pairs — on which N1 is an exact identity (0.2542 observed, 0.2542 null, p = 1.0000) and must not be quoted as a comparator.

These two quantities are one quantity. Every adjacent pair either overlaps or is admissible, so the anatomy-neutral overlap rate is exactly 1 − (admissible / all adjacent): 1 − 839/1,125 = 0.2542 observed, and 1 − 944.6/1,119.1 = 0.1560 under rotation. The surrogates' larger admissible sets and the corpus's overlap excess are the same measurement seen from two sides. We therefore report overlap as the finding, and treat the switch rate as a statistic conditioned on the complement of that finding — which is why it is selection-biased rather than null.

Overlap is large and survives every surrogate that can test it. We report it two ways, because they serve different readers. Scored as the literature does — the share of adjacent *cross-speaker* pairs that overlap — the corpus gives **39.0%**, which is the figure comparable to published values. Scored label-free, as all overlapping adjacent pairs over all adjacent pairs, it gives **0.2542** (286/1,125) against **0.1560** under speaker rotation and **0.1611** under ±2 s jitter, falling to 0.0937 at ±90 s — p = 0.0005 at the permutation floor in every family except N1, which cannot score it. The label-free version is the one to test against surrogates, since no relabelling can move it; the cross-speaker version is the one to compare with other studies. This is the paper's one robust exchange-level result, and it does not depend on speaker attribution being correct.

The alternation and latency statistics cannot be scored on this corpus, and we report that rather than a null. A pair is admissible exactly when it does *not* overlap, so a surrogate that destroys overlap necessarily enlarges the admissible set: 930–972 pairs against the observed 839, which lies 8.6–13.1 standard deviations below every arm's mean and outside all 2,000 draws. The switch rate is therefore scored on sets that differ by precisely the quantity under test, and it correlates with that count in the tight arms and not the wide ones (r = 0.40, 0.39, 0.24, 0.21, 0.14, −0.02, −0.11 from rotation through ±90 s jitter). Regressing each null rate on its own pair count and evaluating at 839 flips the sign of the comparison in **one** of the seven timestamp-moving arms — the ±2 s jitter, which was closest to a tie (0.5378 → 0.4835 against an observed 0.5340); five arms already pointed the same way and rotation does not flip. We report that adjustment as a diagnostic and decline to interpret it, since it extrapolates a decade of standard deviations outside its own support and conditions on a post-treatment quantity. The pre-registered reading of "z ≈ 0 at small jitter" as co-activity is **withdrawn** — not because the statistic came out null, but because it was never scoreable as specified. Detection floors make the same point from the other side — the smallest switch-rate excess any arm could have resolved at α = 0.05 is +2.0 to +2.7 percentage points, so nothing subtler than that was ever in range.

What *is* scoreable is the response window, and it is the section's second result. The registered statistics failed because we operationalised alternation ourselves. Sharma et al. did not leave that open. Their glossary entry for turn-taking reads in full: *"An exchange of codas involving alternating coda production, also called 'adjacent' codas … next-in-sequence codas whose onset occurred within two seconds, but after the termination, of the initial coda."* Both clauses bind — the pair must be cross-speaker **and** inside the 2 s post-offset window. (An earlier version of this section scored the window without the alternation clause; 19 of its 172 numerator pairs were one whale following itself.)

Scored that way (post-hoc; `tools/exp08_response_window.mjs`), turn-taking in this corpus is **far below chance**:

statistic	observed	N4 interval-permutation	N2 rotate	N3 ±2 s
turn-taking rate (cross-speaker responses / all adjacent pairs)	**0.1360** (153/1,125)	0.2337 (**z −6.97**)	0.3121 (z −9.56)	*0.2650 (z −10.71)*
near-simultaneous mass (overlap **or** *any* response)	**0.4071** (458/1,125)	0.4395 (**z −3.99**)	0.4883 (z −7.26)	*0.4898 (z −11.71)*
rate among admissible switches (unstable denominator)	*0.3415* (153/448)	*0.4983*	*0.5716*	*0.5873*

The second row counts responses without a speaker condition (172, not 153), deliberately: it must stay label-free so that relabelling cannot move it, and 286 overlaps + 172 responses = 458. p = 0.0005 at the permutation floor in every arm shown. **We read N4 as the result and treat the jitter arms as upper bounds**, for a reason worth stating: a surrogate must leave same-speaker structure alone, and on the criterion matched to what is scored — the share of same-speaker admissible pairs falling inside the same 2 s window — N4 inflates the observed 0.0486 by 22%, while ±2 s jitter inflates it by **238%** and rotation by 40%. Independent per-coda displacement destroys each whale's own spacing, so the larger z values those arms produce are partly self-inflicted. An earlier draft validated N4 by reproducing the observed same-speaker *median* gap (4.42 s vs 4.42 s); that test is non-discriminative — ±20 s jitter passes it as well — because a median is insensitive to bin mass, which is exactly what this statistic measures.

Denominators are stable across the arms scored: all adjacent pairs number 1,125 observed against 1,119.1 ± 0.4 (rotation), 1,127.0 ± 0.0 (±2 s) and 1,127.0 ± 0.1 (N4, after correcting a branch that had silently dropped five short recordings from every draw and given 1,096.0). The third row's denominator is *not* stable — admissible switches run 448 observed against 610.7 ± 17.8 under rotation — so it carries the same selection bias as the registered switch rate and is reported only for continuity.

Eleven of twelve deployments with at least 20 adjacent pairs fall below their own ±2 s null (sign test p = 0.0063); the dissenter is sw114 at 33 pairs.

The second row matters most: **this is not the overlap result restated.** If it were, adding overlapping pairs back in would erase the depletion. It does not — the combined near-simultaneous mass is *also* below chance. On this corpus whales either overlap, or wait appreciably longer than independence predicts; the band that Sharma et al.'s own definition picks out is depleted at both ends.

Calibration controls. The registered arms carry negative controls (§4); this statistic needed its own, and two were run (`tools/exp08_response_controls.mjs`).

Zero-contingency. Rebuilding whale B's timeline from its own interval multiset, independently of A, gives a corpus whose true cross-whale contingency is zero. Scored against each null, an unbiased null must return z ≈ 0. **N4 does: z = −0.19 ± 1.01. Rotation and ±2 s jitter do not: z = −4.43 ± 0.94 and −2.97 ± 0.82.** A large part of the apparent effect under those two families is the null's own bias rather than signal. Netting each arm against its own zero-contingency baseline leaves the N4 reading essentially unchanged at −6.8; the corrected values for rotation and jitter depend on how the zero-contingency corpus is built and we do not quote them. This is an independent reason to report N4 rather than the arm with the largest z.

Positive control. Injecting genuine responders — moving a fraction *f* of B's codas into the 2 s post-offset window of the nearest preceding A coda — moves z monotonically from −7.04 at *f* = 0 through −3.86, +2.64 and +9.21 to +12.75 at *f* = 1. The statistic detects responders when they are present, so its depletion is a detection rather than a blind spot.

One further asymmetry, which cuts against a simple chorusing reading. Under any symmetric account of joint calling, overlap within a dyad should run about equally in both directions. It does not: **0.4712 (n = 365) one way against 0.3089 (n = 369) the other, p = 0.046** under a per-recording direction-flip null. (An unclustered two-proportion z reads 4.51 and is not the right test: the 734 pairs are nested in 65 recordings, and under the correct null the standard deviation of z is 2.26 rather than 1.) That is the signature a range or focal/non-focal difference would produce — the same channel as the propagation-delay limitation below — and we flag it as unexplained rather than fold it into the interpretation.

The adjacency cap used throughout (30 s) is inert here: no pair of the 1,125 exceeds it, the largest adjacent gap is 26.28 s, and the rate is unchanged at 60 s or uncapped (0.1360 in all three).

A limitation that binds both results in this section. `TsTo` records *arrival* time, not emission time, and the two whales are at different ranges from the recorder. Sharma et al.'s main text describes no propagation correction (we did not examine their Supplementary Information). Injecting a constant per-recording delay τ on the non-focal whale leaves both results intact at $\tau = \pm 0.25$ s and attenuated but far from null at ± 0.5 s; at $\tau = \pm 1.0$ s — about 1.5 km of separation — the response-band depletion and the overlap excess reach their nulls simultaneously. No surrogate can detect this, because every surrogate is applied after the delay is already in the data. Strictly, τ is the *difference* in the two animals' ranges to the recorder, not their separation from each other; the two coincide only when the whales and the recorder are roughly collinear. Both results therefore require that differential range to have been under roughly 750 m on average, which the public deposit cannot establish.

The median switch gap is not resolved, and we make no claim from it. An earlier draft discarded it because "surrogates admit 930–972 pairs against the observed 839"; that reason does not apply to N1, whose admissible count is exactly 839 on every draw. But N1 cannot test latency either, for a stronger reason: because relabelling leaves timestamps untouched, N1's null median gap (4.085 s) is by construction the median over *all* admissible gaps (4.088 s), so comparing the observed switch median (3.616 s) against it only re-states that speaker changes cluster at short gaps — the switch fraction falls monotonically across gap octiles, from 0.92 in the shortest to 0.36 in the longest. That is co-activity, not response. Against every interval-preserving null the sign inverts (N2 z = +11.95; ±2 s +13.00; an inter-click-interval permutation +7.39, validated because it reproduces the observed same-speaker median to z = +0.18). The two families disagree because they ask different questions, and this design does not settle which to believe.

The annotation-artifact check. The serious alternative reading is an annotator splitting one whale's click train into two overlapping codas attributed to two animals. Dual-tag deployments give an independent check using exactly the duplication G0 removes: two tags on two animals record the same encounter from different vantage points, and if overlap were idiosyncratic attribution error on one channel the two tags would disagree. Eight deployments carry two tags; seven support a comparison. Ranked by sample size:

rank	deployment	tag A rate (n)	tag B rate (n)	spread	Fisher *p*	*q* (BH)
1	sw090	0.413 (75)	0.470 (115)	0.056	0.460	0.644
2	sw091	0.345 (58)	0.235 (51)	0.110	0.292	0.644
3	sw133	0.722 (18)	0.368 (87)	0.354	**0.008**	0.058
4	sw078	0.270 (37)	0.429 (7)	0.158	0.404	0.644
5	sw100	0.385 (26)	0.000 (11)	0.385	**0.018**	0.063
6	sw106	0.583 (12)	0.611 (18)	0.028	1.000	1.000
7	sw119	0.333 (9)	0.333 (9)	0.000	1.000	1.000

Two of seven disagree at α = 0.05 (sw133, sw100), both marginal after Benjamini–Hochberg correction. The best-sampled deployment agrees closely and the pooled rate without deduplication is 0.406 against 0.390 after G0, but this check does **not** rule out random attribution error, and an earlier draft of this paper said it did on the strength of three deployments that were ranks 1, 6 and 7 by sample size. Nor does it rule out a *systematic* error — the same pipeline processed both channels. Settling either requires the audio, which cannot currently be joined to these annotations by any public procedure.

5.2 The ornamentation effect appears only at particular segmentation cuts

GAP	Δ observed	naive null (z, p)	stratified null (z, p)
3 s	0.0136	0.0000 (1.3, 0.27)	0.0173 (−0.5, 0.83)
5 s	0.0136	−0.0003 (1.6, 0.15)	0.0090 (0.7, 0.62)
10 s	0.0327	0.0001 (3.1, **0.004**)	0.0130 (2.3, **0.048**)
15 s	0.0408	0.0000 (3.5, **0.004**)	0.0143 (2.7, **0.017**)
30 s	0.0204	−0.0005 (1.8, 0.13)	0.0063 (1.4, 0.23)
8 s *(post-hoc; the paper's own definition)*	**0.0241**	0.0005 (2.4, **0.028**)	0.0114 (1.5, **0.173**)

A sensitivity arm at the authors' own cut. Sharma et al. define a sequence at 8 s; our registered sweep did not include that value, so we added it **post-hoc** (`GAPS = [8]`). At 8 s the positional effect fires against the naive null (Δ = 0.0241, p = 0.028) and does not survive the class-stratified null (p = 0.173). This arm is **a bound, not a non-detection**: its own detection floor is 0.0275, above the Δ it observed, so it could not have resolved the effect it was testing. (Leverage — 382.9 here — is a shuffle-mass exchangeability gate, not a power measure, and we previously misread it as one.) Two caveats bind it further: the glossary's "within 8 seconds" is naturally read onset-to-onset while our GAP is offset-to-onset silence, so the matching cut in our convention may be nearer 7 s; and the same unregistered run produced an S2 rubato result (385 fragments, r = 0.364, p = 0.0005) which we report here rather than omit.

Three further observations. First, the effect appears only under particular cuts — no effect is detected at 3, 5, 8 or 30 s in either arm. **None of those four non-detections is informative**: the smallest Δ each arm could resolve is 0.0327 (3 s), 0.0227 (5 s), 0.0275 (8 s), 0.0299 (10 s), 0.0333 (15 s) and 0.0260 (30 s), and at 3, 5, 8 and 30 s the observed Δ falls below that floor. Only the 10 s and 15 s arms clear their own floors, and both do so narrowly. The cut-dependence argument is therefore weaker than the table suggests: what the sweep establishes is that the effect is resolvable at two cuts and unresolvable at four, not that it is present at two and absent at four. Second, where it fires, the stratified null already produces 40% (10 s) and 35% (15 s) of the observed Δ: class composition accounts for a third to half of the naive effect. Third, and we think most informatively: the paper's ornament flag is operationally defined relative to neighbours **within a ten-second window**, and the only cuts at which the positional effect appears (10–15 s) bracket that definitional window. An effect that lives only at the segmentation scale its own label was constructed at is not evidence of a positional code.

Post-hoc arms (flagged as such; run after the registered result) extend this to the paper's other positional claims. The *initial*-position contrast — the paper's stronger odds ratio — never survives the stratified null at any cut (p = 0.076–0.75). The pooled *edge* contrast fires only at 10/15/30 s (p = 0.0010–0.041) and is absent at 3 and 5 s: the same cut-dependence signature.

We emphasise the scope: this does not test the paper's claim that ornamented codas are rhythmically closer to their temporal neighbours with the final click removed, and a within-class positional excess at intermediate cuts (uncorrected p = 0.017–0.048 at 2 of 5 cuts) is reported, not erased. What fails is

specifically the claim that ornamentation marks sequence edges in a way that survives composition control at more than a privileged segmentation scale.

5.3 Rubato holds under every control we applied

GAP	fragments	r observed	S2 null (z)	S2b null (z)	p (both)
3 s	192	0.106	−0.197 (7.8)	−0.067 (6.4)	**0.0005**
5 s	388	0.265	−0.148 (12.1)	0.021 (10.1)	**0.0005**
10 s	353	0.405	−0.097 (17.5)	0.093 (15.3)	**0.0005**
15 s	324	0.483	−0.079 (20.1)	0.154 (17.6)	**0.0005**
30 s	286	0.522	−0.068 (22.6)	0.173 (19.4)	**0.0005**

p = 0.0005 is the resolution floor at 2,000 iterations; every arm sits on it. The S2b null mean grows with GAP — class composition in time does carry real smoothness, and the control absorbs exactly that — but the observed autocorrelation sits 6.4–19.4 standard deviations above what composition can produce. The inference is identical at all five cuts (only the magnitude grows with window length, which per-fragment centering guarantees mechanically for any genuinely autocorrelated process). Contrast S1, where the *inference itself* flips with the cut.

Robustness (post-hoc, flagged): the effect holds in 58–62% of individual fragments (against a sub-50% baseline from centering bias); inside the majority class alone (z = 5.5–12.0); inside all other single-class fragments pooled (z = 3.7–5.7); and under leave-one-recording-out (worst z = 5.3). Because click count is pinned within class (§2), none of this is click-count variation: successive codas by the same whale drift smoothly in *tempo*, within coda type. That is rubato in the paper's sense, surviving controls that dissolved the ornamentation claim in the same corpus with the same machinery.

6. Relation to the published claims

Paper sub-claim	This work
Adjacent same-type duration drift smaller than chance (rubato a)	**Supported and strengthened** — survives composition + segmentation controls the original test does not apply
Drift direction sustained across sequences (rubato b)	**Consistent** — long strictly-monotone duration runs occur far above chance (e.g. 3/32 length-7 fragments vs ~0.04% expected)
Overlapping whales' durations match: rubato "perceived and controlled" (rubato c)	**Does not survive the missing controls** (experiment 10, pre-registered and adversarially reviewed over five rounds). The covariation is real — r = 0.30 among nearby codas, firing a timeline-rotation null at every window — but a jitter null preserving only the joint activity envelope, with zero cross-whale coupling by construction, reproduces all of it (p = 0.48–0.97). Registered reading: co-activity plus private state, or fine alignment below the co-null's discrimination (the ±2 s jitter re-derives ~90% of pairs identically, a resolution limit registered before freezing). A non-detection of imitation under the controls the published comparison does not apply — not a demonstration that the published effect is artifact. Had a fire survived but failed the overlap-stability check, the registration notes two generators this corpus cannot separate (split-train annotation and genuine chorus-confined imitation); that branch did not occur.
Ornaments mark sequence beginnings (OR 2.00)	**Does not survive** class-stratified control at any segmentation cut
Ornaments mark sequence ends (OR 1.71)	**Cut-dependent**: survives only at 10–15 s, bracketing the flag's own 10 s definitional window; a third to half of the naive effect is class composition
Ornamented codas rhythmically closer to neighbours minus final click	Untested here

6.1 Relation to exchange-prediction work

Two recent results report that the next coda in a Dominica exchange is predictable: Assayag et al. [3] from up to ten preceding codas of the partner whale, and a sequence model of the annotation corpus analysed here [4], whose next-coda perplexity rises 22.7% when input order is shuffled. We have since computed Assayag et al.'s Figure-3 relation in kind on the public deposit (experiment 12, registered and committed at `e4b821d` before the run). It is present: Spearman ρ(CI, CB) = 0.296 on 400 non-overlapping consecutive pairs, Fisher Z = 6.08, p < 1e-9. **The registered gate then failed** — ρ does not exceed its label-shuffle comparator (0.377, p = 0.97) — so by our own registration the verdict is "not reproduced in the public slice", and the registered instrument for the algebraic-bound prediction turned out to be a suppressor. The bound was therefore quantified **post-hoc**. Their two quantities are algebraically coupled: with the first whale's coda beginning at t and lasting d, and its own next coda at t', the coda break is CB = $t\_B - (t + d)$ while the coda interval is CI = $t' - t$, and because the pair is consecutive, $t\_B < t'$. Hence **CB < CI − *d***: the coda break is bounded above by the coda interval minus the first coda's duration, so the two must correlate whatever the whales do. Breaking the CI–CB link while preserving both marginals and that bound yields ρ = 0.248 ± 0.050 — 84% of the observed correlation, with the excess under one standard deviation (one-

sided p = 0.17). By this project's own rule that a non-detection at low power is not written as absence, we do not claim the association is "mostly arithmetic"; we report the bound as a post-hoc structural observation. We also record that degrading the real speaker labels *raises* ρ monotonically (0.296 → 0.372 at 50% flipped). This is a structural observation a reader can check with a pencil; it does not touch their corpus, their models, or their conclusion, and their sample is roughly four times ours with focal attribution we do not have. What §5.1 adds is the missing surrogate discipline for *interpreting* them. Two animals synchronised by a shared activity envelope are mutually predictable by construction; predictability alone cannot distinguish response from co-activity, and on this corpus the co-activity reading is supported on the one statistic this corpus can score cleanly — overlap — and the alternation statistics cannot be scored at all (§5.1). Neither published analysis makes that separation: Assayag et al.'s four null families shuffle feature assignments and coda types across pairs, and none preserves the joint activity envelope while destroying cross-whale coupling; [4] applies no surrogate to exchange timing at all. The one exchange-level statistic our surrogates can score cleanly is overlap — the quantity Assayag et al.'s signed Coda Break metric encodes as its negative range and treats as continuous. Their sign convention shows they met this structure directly, and the algebraic bound above suggests why a correlation of this size need not indicate coupling — their corpus is also not the public deposit analysed here, but a separately annotated 3,933-coda Dtag corpus, so the question has two registrable branches (audit and replication). Two further points of contact. First, [4]'s own input-ablation reports that removing ornamentation and rubato does not significantly affect next-coda prediction — an independent instrument arriving at a conclusion consistent with §5.2's null on the positional claim. Second, what our result adds about the framing is narrower: on this corpus the assumption that whales wait is not supported, and the structure that does survive control runs the other way. A model of coda exchanges compatible with high overlap and envelope-level synchrony — and agnostic about latency, which this corpus cannot resolve — chorusing, in the comparative literature's vocabulary — predicts every exchange-level result reported here, including the predictability that motivated the conversational reading.

A third claim, from outside the CETI group, reaches the opposite conclusion on this same deposit and is reconciled the same way. Bond [7] reports that on `sperm-whale-dialogues.csv` cross-whale "response latency" averages 2.25 s (n = 2,225) against 4.68 s (n = 1,379) for same-whale "continuation", with KS D = 0.56, and reads the 2.08× ratio as consistent with active turn-taking rather than independent vocalisation; he flags the non-independence within sessions and dyads himself and calls for hierarchical confirmation, but applies no surrogate. His released code defines the statistic as the onset-to-onset gap between consecutive rows within a recording, kept when it lies in (0, 30] s and binned by whether the speaker label matches. We reproduce it exactly (`tools/exp08_bond_latency.mjs`, post-hoc: 4.68 s, 2.25 s, ratio 2.08, D = 0.560). His two bins are the corpus-wide adjacent-pair populations of §4 (1,382 same-speaker, 2,239 cross-speaker) under that cap, and they differ by construction. A whale cannot overlap itself, so every same-whale gap is floored at the first coda's own duration (shortest observed 0.83 s; mean coda duration 0.92 s), whereas 901 of his 2,225 cross-whale gaps (40.5%) are pairs in which the second coda begins before the first has ended, with mean gap 0.39 s. Those are the overlaps of §4, entered as short latencies. Removing them leaves 1,324 cross-whale pairs with mean gap 3.52 s, and the ratio falls from 2.08 to 1.33 (D from 0.56 to 0.40). The remainder is not adjudicated by this decomposition. Cross-whale consecutive pairs arise only where two whales are calling, while every stretch in which one whale calls alone contributes same-whale pairs only, and a higher joint calling rate shortens the former without either whale responding to the other. That is the envelope confound the surrogates of §4 exist to remove, and under them the alternation and latency statistics on this corpus are not adjudicated (§5.1). Bond's ratio is therefore the overlap result of §4–§5.1 seen corpus-wide through an onset-to-onset lens, plus a residual this decomposition does not adjudicate — one that co-activity alone could produce and that no surrogate

here bounds; it is consistent with the chorusing reading rather than evidence against it. We note that his analysis is on the same public deposit as ours, that his code is public under MIT, that he anticipated the p-value objection in his own text, and that for the turn-taking and Markov results together he names arousal-dependent call rate as a simpler explanation he could not rule out — the confound this decomposition makes quantitative.

7. What this does and does not show

This is not a refutation of Sharma et al.'s analysis: their tests use definitions, subsetting and modelling choices not reproduced here, and a null on our statistics bears on our statistics. It is evidence about which of the two context-sensitive dimensions denotes structure robust to the two most dangerous degrees of freedom in sequence-level corpus analysis. Rubato is such structure. Ornamentation's positional signature, on this corpus and under these controls, is not established.

The exchange result carries its own scope limits. It licenses nothing about meaning: response contingency is not semantics, and its absence is not the absence of communication — chorusing, overlapping display and simultaneous signalling are communicative in many taxa. What is not supported here is one specific and widely assumed model of what sperm whale exchanges are. It licenses nothing about who responds to whom: speaker indices are per-recording, and the same index in two recordings is not the same animal. And it is timing only — whether a coda resembles what it overlaps is a separate question, one the matching and mirroring literature [3, 5] addresses with different instruments.

Rubato's survival is a statement about structure, not use. Smooth tempo modulation is what respiration, arousal or dive phase would also produce. Experiment 10 — pre-registered, frozen after five adversarial review rounds, with every null validated by negative controls and a synthetic-placebo battery before any statistic ran — tested three provenance signatures. Gap coupling (longer preceding silence, longer coda) is positive at four of five segmentation cuts but fails the registered every-cut criterion. Exchange-mean tempo persists across separations of median 13–37 s but not across the all-≥30 s arm, and the persistence is 65–95% concentrated in one continuous recording scene where a registered qualifier limits the state reading. Cross-whale concurrence — the imitation signature — is fully absorbed by the activity envelope (above). The registered predictions scored one hit and two misses: the corpus certified *less* state signature than the deflationary registration predicted, while still affirming no signal. The honest summary of both experiments: sperm whale rubato is robust structure whose driver and function remain undetermined, and the one published claim of communicative use does not survive independent control.

Tempo, the third dimension, is out of scope: the deposit's tempo labels reconstruct per-coda durations for only 93.3% of rows, so we declined to approximate them. The control granularity equals the authors' own 18-class rhythm inventory; a finer inventory could in principle re-absorb some within-class smoothness as composition — but that inventory is the paper's own, so the claims are tested on their own terms.

8. Reproducibility, data and code availability

All inputs are public. The analysis pipeline is deterministic: fixed seeds, exact permutation enumeration where feasible, and artifacts that reproduce byte-for-byte. Pre-registrations, including the wrong predictions and the corrections they forced, are preserved verbatim in the repository (`experiments/08-turn-taking/`, `experiments/09-rubato-ornamentation/`, `experiments/10-rubato-provenance/`).

Licensing note. The Sharma et al. article is CC BY 4.0, but the data deposit itself (dialogue corpus, rhythm classifications, ornamentation flags) carries no license file. We therefore redistribute none of it: the

repository ships fetch scripts that retrieve every input from the authors' own deposit, plus provenance metadata recording source and alignment validation. All statistics reported here derive from those fetched inputs.

Provenance pin. The deposit is a GitHub repository with no release tags, so its default branch is a moving target: an upstream correction would change every statistic below with nothing to announce it. Both fetch scripts are therefore pinned to upstream commit `7228c8ee` (2024-03-05, the only commit that has ever touched these files) and verify each input's SHA-256 against the exact bytes these results were computed from — on cached files as well as fresh downloads. A mismatch stops the run rather than silently producing different numbers.

```
python3 tools/fetch_corpus.py                          # DominicaCodas.csv + sperm-
        whale-dialogues.csv
./wham/.venv/bin/python tools/fetch_sharma_labels.py  # ornaments.p, rhythms.p; runs
        alignment gate G1
node tools/exp08_turn_taking.mjs        # §5.1
node tools/exp09_rubato_ornament.mjs    # §5.2–5.3 (registered arms)
node tools/exp09_posthoc.mjs            # §5.2–5.3 (post-hoc arms, flagged)
node tools/exp10_rubato_provenance.mjs  # §6 rubato-c row
node tools/exp08_corrections.mjs        # §5.1 dual-tag table, self-overlaps,
                                        #      anatomy-neutral rate, floors
node tools/exp08_response_window.mjs    # §5.1 response window (post-hoc)
node tools/exp08_response_controls.mjs  # §5.1 calibration controls (post-hoc)
node tools/exp12_coda_break.mjs         # §6.1 registered arm
node tools/exp12_controls.mjs           # §6.1 link-broken null (post-hoc)
node tools/exp08_bond_latency.mjs       # §6.1 Bond latency reconciliation (post-
        hoc)
GAPS=8 node tools/exp09_rubato_ornament.mjs   # §5.2 8 s arm (post-hoc)
```

Six of these post-date the pre-registrations they revise and are labelled post-hoc wherever their results appear. The 8 s arm requires the `GAPS` override; without it the tool runs the registered sweep only.

The label fetcher needs NumPy (the class centroids are pickled arrays) and so runs under the project virtualenv; the corpus fetcher is standard library only. Pickles are disassembled with `pickletools` and rejected unless every GLOBAL opcode is a known NumPy constructor, before any load.

References

[1] Sharma, P., Gero, S., Payne, R., Gruber, D. F., Rus, D., Torralba, A. & Andreas, J. Contextual and combinatorial structure in sperm whale vocalisations. *Nat. Commun.* 15, 3617 (2024). doi:10.1038/s41467-024-47221-8 [2] Gero, S., Whitehead, H. & Rendell, L. Individual, unit and vocal clan level identity cues in sperm whale codas. *R. Soc. Open Sci.* 3, 150372 (2016). [3] Assayag, M., Gero, S., Paradise, O. & Diamant, R. Exploring statistical relations between coda sequences of Dominica sperm whale pairs. *Bioacoustics* 35(4), 472–500 (2026). doi:10.1080/09524622.2026.2684665 [4] Sharma, P., Gero, S., Rus, D., Torralba, A. & Andreas, J. WhaleLM: finding structure and information in sperm whale vocalizations and behavior with machine learning. *NeurIPS 2025 Workshop on AI for Animal Communication* (2025); bioRxiv 10.1101/2024.10.31.621071. [5] Gubnitsky, G., Mevorach, Y., Gero, S., Gruber, D. F. & Diamant, R. Automatic detection and annotation of eastern Caribbean sperm whale codas. *Sci. Rep.* 15, 12790 (2025). doi:10.1038/s41598-025-97009-z [6] Schulz, T. M., Whitehead, H., Gero, S. & Rendell, L. Overlapping and matching of codas in vocal interactions between sperm whales: insights into communication function. *Anim. Behav.* 76, 1977–1988 (2008). [7] Bond, A. H. Geometric structure in

sperm whale communication: hyperbolic embeddings, topological analysis, and adversarial robustness. *bioRxiv* (v1, 13 March 2026). doi:10.64898/2026.03.11.711226; code github.com/ahb-sjsu/eris-ketos (MIT).