Does the Markov-Switching Multifractal describe real music? A fitting pipeline and a statistical Turing test

Evan Tabak Atlas
Independent · ORCID 0009-0007-7374-2338
Evanatlas@gmail.com

> Preprint draft (ISMIR-shaped). Figures and numbers below are produced by `run_recovery.py`, `run_corpus_stub.py`, `run_corpus.py` + `scripts/analyze_corpus_maps.py` on the real MAESTRO v3 and Groove MIDI corpora (first external-corpus run 2026-07-12; sha256-pinned downloads; see `out/RESULTS-corpus-v1.md`), and — since 2026-08-27 — `run_competitive.py` for the competitive neural-SOTA arm of §5.1 (`out/RESULTS-competitive-v1.md`). No section is pending data any more; the perceptual protocol (§7) remains design-only.

ABSTRACT

The Markov-Switching Multifractal (MSM) is a parsimonious volatility model – a product of `k` independent two-state switching multipliers – that the Fenophone instrument repurposes as a generative music engine, producing onset streams with measured DFA exponents `alpha ~ 0.70-0.78` and multifractal widths `~ 0.05-0.13`. We ask whether the MSM also *describes* real music. We build a likelihood-based fitting pipeline that treats the MSM as a `2^k`-state hidden Markov model, recover its parameters from synthetic data within ~10% at 10^4 ticks, and pose the closing question as a **statistical Turing test**: resynthesise from the fitted parameters and ask whether a classifier can separate real from surrogate on held-out summary statistics. Fitting 1,264 MAESTRO piano performances and 198 Groove drum takes (beat-time grid, bar covariate), we find a clean dissociation: real piano onset streams are strongly persistent and multifractal (median DFA `alpha = 0.717`, width `0.118` – inside the instrument's own published band), the fitted cascade is genuinely non-flat (`m_hi_eff = 1.62`), and resynthesis retains most of the persistence (`alpha = 0.691`) – while drum-kit onset **counts** are near-white (`alpha = 0.475`) with the cascade structure appearing instead on the **velocity** channel (`vel_m_hi ~ 1.36-1.52` in both corpora). The fitted model is still **not sufficient** at classifier resolution (accuracy 0.758 piano / 0.932 drums vs a ~0.50 permutation null), so MSM-Cox describes real musical intensity at second order without yet being a complete generator of it. A competitive arm (§5.1) places both the engine and the resynthesis on Zhang et al.'s (2025) audio-envelope battery against 381 Billboard hits and three neural generators, domain-matched: the asymmetry we predicted **inverts** – the engine holds spectrum-shape dispersion where the neural systems lose it (`alpha_peak` 1.25x the human IQR against their 0.70-0.86x – **under the IQR ratio only; the comparison reverses under the std and span ratios in the same file**, §5.1) while our arms fall outside the human band on the DFA `alpha` family that all three of them match. A subsequent envelope-domain IAAFT null (§5.1, E8) – the first placed under the width family in

that domain – leaves **5 of 12 shipped packs' `Delta-alpha` indistinguishable from a linear surrogate**.

1. Introduction

The Markov-Switching Multifractal (MSM) of Calvet & Fisher is a parsimonious volatility model – a product of `k` independent two-state switching multipliers – and we read it here as a model of *musical intensity over time*: onset density, velocity, and harmonic tension treated as monotone functions of a single multiplicative cascade. This reading is not hypothetical. The Fenophone instrument already repurposes the MSM as a generative music engine, and its emitted onset streams occupy a measured landscape of DFA exponents `alpha ~ 0.70-0.78` and multifractal widths `~ 0.05-0.13`.

That the MSM can *generate* such streams does not settle whether it *describes* real ones, and that is the research question of this paper: is the MSM a sufficient generative description of real onset/velocity series, or only a convenient engine? We operationalise sufficiency as a Turing test at the level of summary statistics, not perception: fit the model to a real stream, resynthesise from the fitted parameters, and ask whether a classifier can separate real from surrogate on held-out summary statistics. Perceptual sufficiency is a different question with a different method; it gets a companion protocol (section 7), and we collect no human data here.

Our contributions are (i) an exact-likelihood MSM-as-HMM fitting package with three observation models and both canonical and Fenophone parameterizations; (ii) a passing parameter-recovery gate with an explicit identifiability analysis; (iii) the statistical Turing-test harness; (iv) an instrument-self-fit validation and the fitted external corpora; and (v) a competitive placement of both the engine and the resynthesis on a published neural-generator descriptor battery, domain-matched to its audio-envelope domain.

2. Model

The cascade state is the `k`-bit vector of layer states; the process is a hidden Markov chain on `2^k` states whose transition matrix is a Kronecker product of `k` 2x2 matrices.

- **Transition variants.** *Toggle* (Fenophone): a switch always flips, `A_i = [[1-p_i, p_i],[p_i, 1-p_i]]`. *Redraw* (canonical Calvet-Fisher): a switch redraws the layer uniformly, `A_i = [[1-p_i/2, p_i/2],[p_i/2, 1-p_i/2]]`. Both are implemented and compared.

- **Rate ladders.** Fenophone Hz-ladder `p_i = 1 - exp(-f_slow * S^{(i-1)/(k-1)} * dt)` with the engine clamp `f_i <= min(8 Hz, 0.9/dt)` (parameters `(f_slow, S)`); canonical `gamma_i = 1 - (1-gamma_1)^{b^{i-1}}` (parameters `(gamma_1, b)`). Near-reparameterizations, kept so results are quotable in both literatures.
- **Multipliers.** Log-mirrored binomial `M in {m_hi, 1/m_hi}` (primary; geometric mean 1) and mean-1 arithmetic `M in {m0, 2-m0}`. The cascade product is `G = m_hi^{net_state}` with `net_state = #high - #low`.
- **Observation models** (compared by BIC):
  1. **MSM-Cox counts:** `N_t ~ Poisson(lam0 * G_t^{kappa})`, optionally with a per-tick bar-phase multiplicative covariate on `lam0` (meter as an authored seasonal skeleton).
  2. **Velocity volatility:** `log v_t ~ N(mu + (kappa/2) log G_t, sigma^2)`.
  3. **Binary onset indicator:** `P(onset_t) = sigmoid(a + b log G_t)`.

The log-likelihood is exact via the scaled forward algorithm; we validate it against brute-force `(2^k)^T` path enumeration (machine precision, k<=3).

3. Parameter recovery (the gate) – *figure `out/recovery.png`*

Simulating MSM-Cox counts from known parameters (k=4, toggle, Hz-ladder) and refitting over an 8-seed panel at 10^4 ticks recovers, at the **panel median**:

parameter	true	median fit	error
`m_hi`	1.900	1.896	-0.2%
`f_slow`	0.080	0.082	+2.5%
`S`	40.0	38.4	-4.0%
`lam0`	1.200	1.220	+1.6%
`f_fast = f_slow*S`	3.20	3.14	-1.8%

`k` is selected exactly (=4) by both BIC and held-out log-likelihood. **Gate: PASS.**

Identifiability, stated honestly. (a) The Cox count likelihood identifies `m_hi` and `kappa` only through the composite `kappa*log(m_hi)` – a ridge. We therefore fix `kappa=1` for the recovery of `m_hi`, and separately show that with `kappa` free the composite `m_hi^{kappa}` is recovered (2.62 vs true 2.62). (b) Single-realisation slow-rate estimates scatter (sum-of-exponentials ill-conditioning); the reported medians follow the engine's own 8-seed aggregation. Both are reproduced by panel (D) and (A) of the recovery figure.

4. Corpus fits

4.1 Data status – OBTAINED (first run 2026-07-12)

The primary corpora – **MAESTRO v3.0.0** (solo piano) and the **Groove MIDI Dataset v1.0.0** (drum kit) – were downloaded via `scripts/download_maestro.py --confirm` / `download_groove.py --confirm` with sha256 digests pinned on first download (`70470ee2...dde12c`, `651cbc52...1dcc1e`), and preprocessed by `scripts/preprocess_midi.py` exactly as documented: beat-time tick grid at 12 ticks/beat from each file's own tempo map, series trimmed to first->last onset, per-piece gate `>= 2000` ticks with every exclusion logged. Kept: **1,264/1,276** MAESTRO performances (7.0M onsets) and **198/1,150** Groove takes (0.35M onsets; the 952 exclusions are the corpus's deliberately short beats/fills, all `too_short`). Fitting (`run_corpus.py`, checkpointed) uses the first 8,192 ticks, the bar-phase covariate, BIC selection over `k in {5,6}`, `kappa` free, and dt = 1/12 beat so rates read **per beat**. 1,458 pieces fitted; 4 recorded ridge-blowup errors (Poisson-sampler overflow at resynthesis), contained and listed in `out/corpus_fits.csv`. (Lakh MIDI remains a future robustness subset.)

The two MAESTRO counts, reconciled (2026-08-25). Both appear below and they mean different things: **1,264** is the corpus *kept* count (1,264 of 1,276 pass the preprocessing gate), while every corpus statistic in §4.3 and §5 – the medians `m_hi_eff` 1.619, `alpha` 0.717, width 0.118, and the Turing-test row – is computed over the **1,260** performances with a *usable fit*, the 4 errored rows above excluded. A headline attributed to "1,264 performances" is naming the corpus; a median quoted at `n = 1,260` is naming the fitted set. The Groove side has no such split (198 kept, 198 fitted).

4.2 Instrument-self-fit stand-in – *figure `out/instrument_fits.png`*

As the available dataset we fit engine streams (four scaffolds, rendered via `foundry` at 8192 ticks). Fitting the MSM-Cox per stream and BIC-selecting `k` gives (from `out/instrument_fits.csv`):

scaffold	k	m_hi_eff = m_hi^kappa	f_slow	S	lam0
launch-beat-driven	5	1.005	~0	~1	1.75
launch-cinematic	5	1.065	0.053	~1	1.56
launch-warm-ambient	5	1.047	0.044	~1	1.16
prototype-beat-forward	5	1.000	~0	5.58	0.45

BIC selects k=5 over k=6 for every stream. The *fitted* cascade is near-flat (`m_hi_eff ~ 1`), with `lam0` tracking each scaffold's onset density (0.45-1.75). This is a real and important result, not a bug: the maximum-likelihood MSM-

Cox **under-weights the long memory that DFA measures**. The engine's emitted onset counts have `alpha ~ 0.80` (genuine long-range correlation from the slow cascade layers), but their contribution to the per-tick Poisson likelihood is small relative to the marginal count distribution, so the MLE flattens the cascade toward near-Poisson. (Verified directly: the structured cascade `m_hi=1.8, f_slow=0.1, S=40` scores log-lik -10754 against the near-flat MLE's -7364.9.) This is the well-known reason the MSM econometrics literature estimates by GMM / spectral methods rather than plain ML, and it is exactly the "fitted observation model misspecified against the generating channel" case the recovery section previews. The fits are explicitly instrument-self-fit, not a claim about human corpora.

4.3 External-corpus fits – *figures `out/corpus_param_maps.png`, `out/corpus_alpha.png`; full report `out/RESULTS-corpus-v1.md`*

The corpus run **overturns the section-4.2 expectation**: on real piano the ML fit does *not* collapse. Medians over the fitted pieces (rates per beat; `out/corpus_fits.csv`):

corpus	n	k	`m_hi_eff`	`f_slow`/beat	`lam0`/tick	real `alpha` (shuffled)	real width	Fano
MAESTRO (piano)	1,260	5	**1.619**	0.0019	0.233	**0.717** (0.500)	**0.118**	1.552
Groove (drum kit)	198	5	**1.027**	0.0012	0.421	**0.475** (0.503)	-0.031	1.157

Three findings:

1. **The multifractal-intensity premise holds for the piano idiom.** Bar-adjusted onset density in real performances is strongly persistent (83.2% of pieces `alpha > 0.6`; IQR 0.634-0.775) with positive multifractal width – quantitatively **inside the engine's own published band** (`alpha ~ 0.695-0.770`, width `~ 0.054-0.126`). This is the first non-self-referential support for the programme's core premise.
2. **The ML collapse was the channel's, not the estimator's.** Piano counts are overdispersed (Fano ~ 1.55), so the likelihood rewards a structured cascade: `m_hi_eff` median 1.619 with 92.6% of pieces > 1.2, and the resynthesis retains most of the persistence (synth `alpha` 0.691 vs real 0.717). The engine's emitted counts, by contrast, present near-Poisson marginals that mask their long memory – explaining section 4.2 without excusing it. (The kappa-log(m*hi) ridge and the slow-rate ladder remain per-piece-unstable; `S` and `kappa` are quotable only as panel medians.) **Null-floor calibration (2026-08-26, re-cut 2026-08-30 as `out/RESULTS-nullfloor-v2.md`; v1 is superseded and retained):** because the likelihood is rewarded by the marginal overdispersion, `m_hi_eff` carries a large **zero-correlation floor**. Two arms measure it, and they are on **different corpora** – a distinction this paragraph previously elided. **(a)**

The real-data decorrelation arm is GiantMIDI, not MAESTRO. On **56 audio-transcribed GiantMIDI performances** – a different corpus at a different marginal (Fano median **1.797**, mean **0.379**/tick, against MAESTRO's 1.552 and 0.405) – the data's own shuffle and an iid negative-binomial null matched to each piece's own (mean, Fano) both fit **≈ 1.76–1.81** against real **1.80**, so destroying every temporal correlation costs nothing measurable. **No shuffle or Fano-matched-negbin arm has ever been run on the 1,260 MAESTRO pieces**; `run_corpus.py` shuffles only for a DFA-α null and `out/corpus_fits.csv` carries no shuffled or negbin `m_hi_eff` column. That direct null is queued (`research/TODO.md`) and is the one this claim actually wants. **(b) At MAESTRO's own operating point the floor is synthetic**, from the iid-negbin curve: mean- and Fano-matched to the corpus marginal (mean **0.4046**/tick, Fano 1.552, k = 5) it fits m̂_hi ≈ **1.605** (λ²_floor ≈ **0.708**), i.e. the sealed 1.619 sits **at** its own zero-correlation floor. *(v1 published 1.794 / 1.690 for this point. Those numbers were generated at the fitted **latent** Cox rate $\lambda_0 = 0.233$ where the generator takes a **marginal** mean, so the whole curve sat at 0.576× the corpus's operating point; v2 corrects the parameterization. The GiantMIDI arm is unaffected and reproduces to numerical identity on every committed median (0.0); one weakly-identified kappa cell (piece p020, negbin row) differs in the third decimal.)_ A Poisson control (Fano ≈ 1) reads 1.00 and a notation-only null overshoots off the crosswalk's domain, so the estimator is sound and its floor is set by overdispersion. So `m_hi_eff = 1.619` is a **marginal-overdispersion** statistic (not, on its own, evidence of temporal cascade structure), and the Feno-program §4 crosswalk of it ($\lambda^2 \approx 0.68$) is a discrete crosswalk with a **now-calibrated** null floor – consistent with the MFDFA width **0.118** sitting in the continuous band rather than at 0.68, **once the unit change is stated**: for a log-normal cascade $\zeta(q) = qH - \lambda^2 q(q-1)/2$ gives $h(-2) - h(2) = 2\lambda^2$, so a two-point width of 0.118 implies **$\lambda^2 \approx 0.059$**, not 0.118 — inside the continuous band 0.03–0.12 (`research/README.md`'s width-disambiguation table carries the pair). A further caveat on any λ^2 read off this estimator: the §4 crosswalk is defined only on $\hat{m}_{hi} \in (1, 2)$, and **264 of the 1,260 MAESTRO fits (21 %)** and **16 of the 56 GiantMIDI fits (29 %)** sit at or above 2, off its domain, with the headline λ^2 taken as the crosswalk of the median rather than the median of the crosswalks. The temporal-structure claim rests on the DFA `alpha 0.717` vs shuffled `0.500` dissociation (finding 1), which the floor leaves intact.

3. **Drums dissociate.** Bar-detrended drum-kit **counts** are near-white at this grid (`alpha ~ 0.475 ~` shuffled) and the fitted cascade is flat – but the **velocity** channel is not: on the every-8th-piece velocity-fit subset (n=182),

`vel_m_hi ~ 1.52` (piano) and `~ 1.36` (drums) at `kappa ~ 0.5`. In the drum idiom the multiplicative structure lives in loudness (and, per the microtiming literature, in timing deviations), not in beat-grid onset density.

5. The statistical Turing test – *figures `out/turing_test.png`, `out/turing_classifier.png`*

For each stream we resynthesise a matched surrogate from the fitted parameters and compare, on bar-adjusted onset series: DFA `alpha`, MFDFA width, and the count-dispersion (Fano) profile. We then pool windowed summary statistics (mean, variance, Fano, activity, lag-1 autocorrelation, local DFA `alpha`) from real and surrogate windows and train a from-scratch logistic classifier to separate them, scoring held-out accuracy against chance (0.5) and a label-permutation null. **A strong classifier that cannot beat its permutation null is the sufficiency verdict.**

We validate the test in **both directions**:

operating point	real `alpha`	resynth `alpha`	classifier acc (chance 0.5)	perm-null p95	verdict
positive control (true MSM-Cox)	0.79	0.81	**0.474**	0.59	**SUFFICIENT**
engine onset streams (fitted Cox)	0.80	0.54	**0.987**	0.56	**INSUFFICIENT**
MAESTRO piano (n=1,260)	0.717	0.691	**0.758** (69,540 windows)	0.500	**INSUFFICIENT**
Groove drum kit (n=198)	0.475	0.511	**0.932** (5,806 windows)	0.510	**INSUFFICIENT**

The **positive control** (fit + resynthesise a genuinely MSM-Cox stream) reproduces the DFA `alpha` and leaves the classifier at chance – so the machinery correctly calls a *true* model sufficient, proving it is not rigged to pass or fail. The **engine streams** are the honest negative: the near-flat MLE resynthesis drops from `alpha ~ 0.80` to `~ 0.54` (near-white), and the classifier separates real from surrogate at 0.99 accuracy.

The **external-corpus verdicts** land in between, and differently per idiom. On piano the fitted model is a far better generator than it was for engine streams: the resynthesis keeps the persistence (0.691 vs 0.717) and the classifier, despite 12x more windows, drops to 0.758 – yet that is still decisively above its permutation null, so MSM-Cox-as-ML-fitted describes real piano intensity **at second order without being a sufficient generator** of it (the residual separability rides on higher-order window

features, e.g. the dispersion profile). On drums the near-flat fit fails harder (0.932; the resynthesis *over*-disperses, synth Fano 1.47 vs real 1.16). Both verdicts keep the original motivation: a long-memory-aware estimator (GMM/Whittle) and the velocity/combined channel are the next estimators to run – the velocity subset already shows non-flat structure in both corpora.

5.1 The competitive Turing test (neural-SOTA arm) – *run 2026-08-27; full report `out/RESULTS-competitive-v1.md`*

The sufficiency verdict above asks whether *our* fitted model can be told from real music. The competitive question – how the MSM/Fenophone compares to the current state of the art in generative music – became directly testable when Zhang, Sun & Xiong (2025, ACM MM '25; public code + generated corpus at `github.com/zhangkkevin/billboard-ai-fractal-comparison`) applied essentially this paper's descriptor battery (DFA `alpha`, singularity-spectrum width, `alpha`-peak, skewness, the generalized- Hurst ladder h(q), and the Jensen-Shannon divergence between f(`alpha`) spectra) to 381 Billboard hits (1950-2024) against matched Suno, DiffRhythm, and YuE outputs, and found the neural generators systematically **under-disperse** the human multifractal signature – the songs are too similar to each other in fine-scale scaling, rather than the average song being wrong. We score three arms of our own on that pinned battery: 96 engine streams (12 shipped scaffolds x 8 seeds, all at the shipped cascade point), 8 renders of the MAESTRO median fit, and 24 per-piece resyntheses. The domain caveat is closed by construction, not argued away: every stream is reduced to a 150 Hz amplitude envelope through their exact Hilbert -> 25 ms -> 150 Hz path and scored on their q-grid (so `h(q)` is quoted at the real rungs ±2.195122, never relabelled "q = 2"), at 217 s – the Billboard median duration – so all 128 streams get the identical 36-rung DFA window ladder the median-length human song gets. Their published per-song numbers reproduce exactly from our port first: 1,490 rows, max |Δ`alpha`| 1.55e-15, zero rows off by more than 1e-6.

The asymmetric hypothesis inverts. Medians against the human band (Billboard p05-p95), with dispersion as the IQR ratio to Billboard – the axis the published deficit actually lives on:

descriptor	human band	Billboard	neural IQR ratios	engine median (n=96)	engine IQR ratio
DFA `alpha`	0.838 - 1.164	0.988	0.95 - 1.28	**1.263** (outside)	0.40
`alpha_width`	0.736 - 2.753	1.656	0.47 - 0.81	1.323	0.48
`alpha_peak`	1.155 - 2.322	1.329	0.70 - 0.86	1.605	**1.25**
`spectrum_skew`	-0.723 - 0.765	0.532	0.58 - 0.69	0.260	**0.86**
`h(-10)`	1.564 - 3.438	2.467	0.41 - 0.86	2.354	0.47
`h(-2.195)`	1.585 - 3.166	2.295	0.47 - 1.05	2.081	0.48
`h(+2.195)`	1.006 - 1.365	1.160	0.88 - 1.22	**1.407** (outside)	0.63
`h(+10)`	0.851 - 1.206	1.004	0.78 - 1.24	**1.208** (outside)	0.36

Read in the direction the hypothesis predicted, this is a swap. The engine arm *is* inside the band across the whole spectrum-shape family (94-100 % per-song membership on `alpha_width`, `alpha_peak`, `spectrum_skew`, `h(-10)`, `h(-2.195)`; so are the resynth arms, bar `msm-resynth-pieces` falling just under the band on `h(-2.195)`), and it out-disperses all three neural systems on two of those five – `alpha_peak` at 1.25x the Billboard IQR against their 0.70-0.86, `spectrum_skew` at 0.86x against 0.58-0.69 – **under the IQR ratio, and only under the IQR ratio** (see the two scope notes below).

Scope of the dispersion win (dated note, 2026-08-30). `out/competitive_bands.csv` computes three dispersion ratios per cell, and the win holds under one of them. Under the **std ratio** Suno out-disperses the engine on both descriptors (`alpha_peak` 0.900 vs 0.533; `spectrum_skew` 0.761 vs 0.508) and so does YuE (0.816 vs 0.533; 0.519 vs 0.508); under the **span ratio** Suno does again (1.029 vs 0.827; 0.944 vs 0.519). The IQR is the defensible statistic here, and the reason is measurable rather than rhetorical – Billboard's `alpha_peak` has skew +3.14, excess kurtosis +12.5 and std/IQR 2.55, with 40 % of its standard deviation coming from the outer 5 % of songs, and `alpha_peak` is `h(q)[argmax f(alpha)]`, so its tail is populated by argmax flips rather than by spread – but that rationale was written after the statistic was seen to win, which is exactly why the statistic is now being **pre-registered before the pending full run** (`research/TODO.md`). The honest form: the win is real under a robust spread statistic, it is not robust to the choice of statistic, and `run_competitive.py`'s `dispersion_ok` / `closes_neural_gap` verdicts are IQR-ratio verdicts. Full working: `out/RESULTS-competitive-v1.md`, "Dated note (2026-08-30)".

Cross-container portability caveat on these two descriptors (dated note, 2026-08-30). `out/RESULTS-competitive-v3.md` found, on **byte-identical audio** (`wav_sha256` reproduced), that the **negative-q branch of the MFDFA family is not portable between containers**: `dfa_alpha` and the positive-q rungs agree to 1e-16 and 1e-8, but Delta-alpha, `alpha_peak`, `spectrum_skew`,

h(-2.195) and h(-10) drift by order 1e-2, and up to **0.34 on `alpha_peak`** and **0.36 on `spectrum_skew`** where the argmax flips – against Billboard IQRs of 0.207 and 0.319. Every `alpha_peak` / `spectrum_skew` number in this section, the dispersion ratios above included, is quoted **within one container** and carries that error when compared across runs. On the other three (0.47-0.48x) it sits beside DiffRhythm's 0.41-0.47x rather than Suno's or YuE's 0.75-1.05x. But on the positive-q/DFA family, where all three neural systems sit inside the band with Billboard-matching spread, **our medians sit outside it** (per-song membership 8 %, 29 %, 48 %). Band membership itself decides nothing here and we do not lean on it: all three neural systems place 74-100 % of their songs inside the human band on every descriptor, because the 1950-2024 band is wide and the generators are narrow.

The same inversion shows up in the pooled-spectrum divergence, once it is given a floor. Protocol-A JSD is strongly sample-size dependent under their binning, so each system is scored against a Billboard-vs-Billboard null at its own `n`: Suno 0.199 vs a floor of 0.119 (excess **+0.079**), YuE 0.310 vs 0.132 (**+0.179**), DiffRhythm 0.393 vs 0.119 (**+0.274**), and our engine 0.709 vs 0.401 (**+0.309**) – the largest excess in the panel. **Two caveats travel with that ordering (dated note, 2026-08-30).** (i) *Design mismatch:* the statistic is a 381-vs-`n` comparison while the floor is drawn `n`-vs-`n`, so the floor carries sampling noise on both sides where the statistic carries it on one – it is not the null of the quantity it is subtracted from, and the design-matched floor (`billboard[381]` vs `billboard[n]`) has not been run. (ii) *Bounded scale:* protocol-A JSD is bounded in [0, 1] and saturating, so an **additive** excess is not commensurable across floors that differ by 3.4x – the maximum attainable excess is 0.599 at the engine's floor and 0.881 at Suno's and DiffRhythm's. The **scale-free contrast**, excess / (1 - floor), computed from the committed `out/competitive_jsd.csv`, preserves the ordering and widens it: **engine 0.515

> DiffRhythm 0.311 > YuE 0.206 > Suno 0.090**. Quote it beside the raw excess; the conclusion is better supported by it than by the additive number as published. The two resynth arms come out *at* their floors (0.809 against 0.863 at `n` = 8; 0.690 against 0.687 at `n` = 24, both inside the floor's own p05-p95), but those floors are so high on a statistic bounded by 1 that "at the floor" is nearly unfalsifiable at those panel sizes; we report them and decline to bank them.

The obvious mechanism for the `alpha` excess is falsified. At the shipped default the slowest cascade layer runs at `f_slow` ~ 0.10 Hz, inside the band of timescales the DFA ladder integrates over, so we swept the Change Rate dial

(4 scaffolds x 2 seeds x 6 rungs, everything else at defaults) with the pre-registered prediction that `alpha` would be monotone in it. It is not: the pooled median moves non-monotonically (Spearman rho = -0.31) with a total swing of **0.084** across a 50x range of `f_slow` (0.010-0.500 Hz), and stays outside the human band at all six rungs. There is nowhere on that dial where the engine's persistence exponent becomes human. What the sweep does show is that the descriptors are not one effect: at dial 0 the `h(+2.195)` in-band share rises to 88 % (from 12 % at the default) while `alpha` barely moves and `alpha_width` moves *away* from the human median. Variance decomposition points elsewhere – 59 % of the sweep's `alpha` variance is between scaffolds, against 26 % across the dial – and the 12 scaffold medians order against tempo (rho = -0.54, p = 0.07) and onset density (rho = -0.63, p = 0.03), with exactly one scaffold (`base-terracotta`, 1.109) inside the band. That is a hypothesis about the authored note process, explicitly not a claim.

The standing limitations are structural and are stated in full in the run report: our "corpus" is 12 shipped presets x 8 seeds with 66-95 % of descriptor variance between presets (a catalogue, not 75 years of practice), `msm-resynth-median` measures seed jitter rather than dispersion, the resynth arms render at the *nearest reachable* operating point (88 % of MAESTRO fits rail Spread and 76 % rail Change Rate – the section-6 ridge, now measured against the instrument's reachable set), the genres do not match (Billboard pop mixes against an instrument catalogue; `base-neo-soul` stands in for piano), the neural-miss rule is calibrated from the baselines themselves, and 51 of 96 engine streams are hard-clipped pre-envelope – an effect that collapses to |rho| <= 0.20 under scaffold-centring, leaving a level-trimmed re-render as the decisive unrun test. **This is a descriptor placement and nothing more:** no perceptual quality or authenticity is inferred from any gap above, in either direction, which is the same limitation Zhang et al. flag about their own comparison. Full record, tables and figures: `out/RESULTS-competitive-v1.md`.

v2 (same day): the four registered follow-ups ran, and the picture sharpens in both directions. The level-trimmed re-render turned out not to be expressible in the scaffold contract — the schema exposes no pure-gain field for the synth primitives (2 of the 31 voices across the seven clipping scaffolds are samplers, the only primitive with a level), so that test now waits on an engine-side gain flag and v1's dispersion wins stay qualified rather than certified. The note-process manipulation (2 bases x 6 axes x 43 loader-validated variant rungs x 4 seeds, 180 streams, 0 failures) confirms the tempo/density hypothesis in direction and falsifies it in magnitude — monotone (rho -0.50 to -0.89) but worth only 0.2-0.5 Billboard IQRs across 2.7x/5x ranges, with the room null on `alpha` across a 15x range of T60 —

while the axis included as a control dominates: **per-note decay moves `alpha` by 1.57 Billboard IQRs (rho = +0.981) with `notes.csv` bit-identical**, and decay x0.25 moves `base-terracotta`'s DFA `alpha` **median** from 1.110 to **0.986**, onto Billboard's 0.988 – a 0.88 Billboard-IQR displacement. **(Corrected per `ADDENDUM-competitive-v2.md` A1: that is a median-displacement finding, not a membership one.** v2 read the same rung as putting "all eight descriptors inside the human band with no membership trade-off"; the manipulation's contribution to membership is **zero**, because the **shipped** `base-terracotta` is already 8 of 8 descriptors in band at 8 of 8 seeds on the v1 panel, with no variant at all. Any per-pack CI gate built on that membership would be built on a criterion the shipped pack already satisfies, so the v2 owner recommendation proposing one is withdrawn with the reading.) Consistently, an audio-only piano-matched carrier (event layer held bit-identical, asserted by sha256) erases both v1 resynth band failures — the resynth DFA-`alpha` excess was the neo-soul carrier, not the fitted cascade — yet the engine arm at the shipped point stays outside at 0 % membership. And re-fitting all seven systems on the one DFA ladder common to every song (27 rungs, top 3.66 s) shows the engine's `alpha` excess is **statistically indistinguishable between the full and common ladders** (+0.100 → +0.094; ratio 0.97, 95 % CI [0.75, 1.27] over 4,000 bootstrap resamples, 39 % of them putting the ratio above 1) **(dated note, 2026-08-30: v2 quoted this as "94.7 % survives" to three significant figures with no interval. The load-bearing result is not the ratio but the excess itself, which is > 0 in 4,000/4,000 resamples on both ladders – full 95 % CI [0.068, 0.125], common [0.065, 0.118])**, while the two duration-mismatched published gaps move materially (Billboard−DiffRhythm triples to +0.052; Billboard−YuE shrinks 42 % to −0.047) and the resynth arms cross into the band only because the band itself rises at short scales past their scale-flat `alpha`. Net: the persistence-exponent failure is signal, and its causal locus is per-note energy decay in the audio layer, not the cascade dials or the authored note process. Full v2 record, updated verdict table and the separated owner recommendations: `out/RESULTS-competitive-v2.md`.

v3, E7 and E8 (2026-08-27/28; propagated here 2026-08-30): the ladder crossed, the catalogue widened, and the first null placed under the width family in this section's own domain. Three records landed after the v2 paragraph above was written, and all three qualify it.

- **v3 (`out/RESULTS-competitive-v3.md`) – the carrier decay ladder (E5) and mediation-blocking density at deep decay (E6), both pre-registered.** E5's registered prediction held: the carrier's DFA `alpha` median crosses Billboard's band top at **x0.387** (log-interpolated; x0.416 fitted), inside the registered x0.1-x0.5, with the carrier's dose-response

1.51x steeper than the `base-terracotta` fit it was transported from. E6 returned the registered **INTERMEDIATE** reading – the pooled `alpha` swing falls from 0.0586 at x1 decay to 0.0415 at x0.0625, **70.7 %** retained – so per the registration **no mediator is named**. v3 also carries the cross-container portability finding stated above, which is the reason every `alpha_peak` / `spectrum_skew` number in this section is now quoted with a container caveat.

- **E7 (`out/RESULTS-voicesets-v1.md`) – the 14 shipped voice sets, measured for the first time, exploratory by registration (B5) with no prediction attached.** Only **2 of 26** configurations sit inside Billboard's band on all eight descriptors at all 8 seeds; one is `base-terracotta`'s roster (already known), the other is **`launch-cinematic/mallet`, which already ships** and whose DFA `alpha` median of **1.0658** against its roster's 1.2585 is a **1.37 Billboard-IQR** move carrying `alpha` from 0/8 streams in band to 8/8. Across the catalogue there are 17 (set, descriptor) membership changes and 11 median band-edge crossings, nine of them inward, and the **post-hoc** (explicitly not registered) expectation that modal-bodied sets sit lower on `alpha` held on 11 of 11 such sets, Spearman rho = -0.961. The sets **add** between-configuration spread: on v1's own per-stream statistic `alpha_peak`'s IQR ratio rises from 1.248x (96 roster streams, reproducing v1 exactly) to **1.351x** (208 streams) and `spectrum_skew`'s from 0.858x to **0.925x** – which widens the two wins above without changing their IQR-only scope.
- **E8 (`out/RESULTS-envelope-null-pilot-v1.md`) – the envelope-domain IAAFT surrogate null, pilot, registered (B4) with either direction acceptable. It is the first null ever placed under the width family in the domain this section's descriptors live in, and it is mixed: 7 of 12 shipped packs' envelope Delta-alpha exceeds their own IAAFT null's p95, and 5 of 12 do not** (`base-deep-techno`, `base-lofi-tape`, `base-neo-soul`, `launch-warm-ambient`, `premium-hollow-tension`). For those five the multifractal width of the audio envelope is not distinguishable from what linear correlation plus the amplitude distribution alone produce. The excess that is there is **asymmetric across the spectrum**: `alpha_peak` exceeds its null on 12/12 packs (median z = 17.7) and h(-2.195) on 10/12, while h(+2.195) exceeds on 3/12 and h(+10) on 1/12, with 11/12 packs at or below their own null on h(+10).

  Two disambiguations are mandatory wherever E8 is quoted. (i) *Which width.* E8 measures **Delta-alpha, the Legendre spectrum width of the audio envelope** – this section's statistic – and **not** the two-point MFDFA width `h(-2) - h(2)` of event-time onset streams that the engine's own IAAFT gate and `research/data/nulls-v*.csv` score. The two share a

name and nothing else; their numeric ranges differ by roughly 10x, and a number from one is meaningless against the other's nulls (`research/README.md`, "Width means two different statistics here"). The catalogue-wide 11/11 IAAFT result quoted elsewhere in the program is the **onset-domain** statistic and is untouched by E8. (ii) *The halving.* IAAFT truncates rather than pads to a power of two, so every 32.6k-sample envelope becomes **16,384 samples (109.2 s)** – registered before the run, in B4, precisely so it could not be discovered afterwards. **No E8 number is comparable to the Billboard band or to any descriptor CSV in v1/v2/v3**, and the bound it places is on the truncated series. It is still the first null those packs' width has been set beside in its own domain, and five of them do not clear it. E8 also runs 12 packs x 7 descriptors x 2 tails = 168 one-sided rank tests at 19 surrogates, whose floor p is 1/20 = 0.05 with no multiplicity control, and its seven descriptors are strongly dependent (all read off one H(q) curve; `alpha_peak`, `spectrum_skew` and Delta-alpha are algebraic functions of the same 42-point ladder): about 8.4 of 168 tests are expected at the floor under a global null, so the headline counts (7/12, 12/12, 11/12) stand well clear of chance while the weak per-descriptor counts (h(+2.195) at 3/12 upper and 4/12 lower) do not and must not be read as substantive. The scheduled 128-stream run moves to 99 surrogates, which fixes the resolution half.

6. Interpretation targets, answered – *`out/corpus_group_maps.csv`, `out/corpus_separability.csv`*

- **Does fitted `k` track formal/textural depth?** Weakly at best: BIC prefers `k=5` for ~63% of pieces in *both* corpora (k=6: 37.4% piano, 35.4% drums); k-selection is not idiom-diagnostic at this length/grid.
- **Does `m_hi` track expressive dynamics?** As a composer-level gradient, yes: Bach sits lowest (`m_hi_eff` 1.454, `alpha` 0.604, width 0.067 – motor-rhythmic, terraced), Beethoven/Liszt/Schumann highest on `alpha` (0.775-0.783), Debussy widest (0.206), Chopin/Schubert/Liszt at `m_hi_eff ~ 1.67-1.73`.
- **Do drum corpora sit at different `(f_slow, S)` than piano?** The stronger, cleaner statement is that they sit at a different `m_hi_eff` entirely (1.027 vs 1.619): the count channel carries no cascade for drums (section 4.3). The rate-ladder axes are ridge-unstable per piece and are not the quotable coordinates.
- **Composer/genre separability.** MAESTRO, top-8 composers (n=953), one-vs-rest logistic CV: permutation chance 0.190 -> MSM parameters

0.276, (`alpha`, width) baseline 0.337, **combined 0.374**. So composer identity *is* decodable from these statistics, the spectral pair is the stronger single descriptor, and the MSM parameters add ~4 pp on top of it – the honest answer to "does the MSM carry more than `alpha`?" is "some, jointly, not alone." Groove styles (rock/funk/afrobeat, n=73): nothing beats chance (0.540 vs 0.500) at this sample size.

7. Perceptual protocol (design only)

No human data. The design: render fitted parameters through `foundry render` at the nearest engine operating point and run a 2AFC "which is the model" study. Out of scope for the statistical sufficiency claim here.

8. Limitations (honest)

- **Two idioms, one run.** The corpus results cover solo classical piano (MAESTRO) and studio drum kit (Groove) – claims are bounded to those idioms, not "music". The analysis-defining choices (12 ticks/beat, the 8,192-tick cap, bar-template handling) are the documented protocol's, applied uniformly, but not yet sensitivity-swept.
- **Channel misspecification.** MIDI is not audio; the engine's latent->emitted channel (and a real instrument's) is not the fitted observation model. The Fenophone's own channel study (`research/papers/04-quantization-channel.md`) is the companion.
- **Grid/tempo choices are analysis-defining.** Beat-time quantization at 12 ticks/beat, the `>=2000`-tick segmentation, and bar-template handling all shape the fits; documented and logged, not silent.
- **Cox ridge.** `m_hi` and `kappa` are jointly identified only through their product from counts alone; velocity or a combined channel is needed to separate them.
- **Slow-rate variance.** Slow layers are weakly identified at feasible lengths; we report panel medians and CIs, not single-series point estimates.
- **MFDFA width is high-variance.** The headline MAESTRO width (+0.118) has a wide, right-skewed IQR (0.026-0.208) at this length. The companion Cox-vs-Hawkes paper (`../msm-cox-hawkes/`) reports a *negative* median width (-0.069) for MAESTRO, computed on a 120-piece seeded subsample of dequantized event times with a different estimator; that is the same statistic's small-subsample, different-pipeline reading of the same heavy-tailed distribution, not a contradiction (see that paper's §7 reconciliation note). The full-corpus onset-density +0.118 here is the program's canonical width figure.

9. Reproducibility

Everything is seeded. `pytest` (estimator anchors + forward-vs-brute-force + recovery sanity), `run_recovery.py` (the gate), and `run_corpus_stub.py` (self-fit + Turing test) run offline except the engine render step; `scripts/render_competitive_streams.py` and `run_competitive.py` (§5.1) likewise need only the engine once the baseline CSVs are on disk. See README.md for commands, runtimes, and checksums.

Data provenance. MAESTRO v3.0.0 (Hawthorne et al. 2019) is distributed under CC BY-NC-SA 4.0 and the Groove MIDI Dataset v1.0.0 (Gillick et al. 2019) under CC BY 4.0. Neither corpus is redistributed by this work: only derived quantities – fitted parameters, DFA/MFDFA statistics, and the figures built from them – are committed and deposited. The raw corpora are fetched at analysis time by the sha256-pinning download scripts (`scripts/download_maestro.py`, `download_groove.py`), which verify each archive against its recorded digest before use; see `research/data/PROVENANCE.md` for the full notice. The §5.1 baselines are Zhang, Sun & Xiong's (2025) published per-song CSVs at commit `51b243f4`, fetched and sha256-verified per run by `scripts/download_zhang_baselines.py` and **not redistributed** (their repository carries no LICENSE at that commit); what is committed here is the derived summary, with attribution.

References

Author-year; keys map to the shared bibliography `../../papers/references.bib`. Companion papers are the Fenophone research series (Atlas 2026a-d).

Companion records of the Feno program (DOIs reserved 2026-08-30; published 2026-08-31). Each was staged as a private Zenodo draft with a reserved DOI and published 2026-08-31 — every DOI below is live and resolves. The deposit map is `../../zenodo/records.json`, and because the source repositories are private every record ships its own reproduction materials.

- Atlas, E. T. 2026a. "The Fenophone: A Musical Instrument Built on the Markov-Switching Multifractal, with CI-Verified Spectral Invariants." Fenophone research series, paper 01. Zenodo DOI 10.5281/zenodo.22180423 (published 2026-08-31).
- Atlas, E. T. 2026c. "The Markov-Switching Multifractal as a Generative Model of Musical Intensity: An Instrument, Measurements, and a Research Programme." Fenophone research series, paper 03. Zenodo DOI 10.5281/zenodo.22180429 (published 2026-08-31).
- Atlas, E. T. 2026. "msm-corpus-fit: MSM Fitting Pipeline, Statistical Turing Test, and Committed Corpus Results." Software record supplementing this paper. Zenodo DOI 10.5281/zenodo.22180443 (published 2026-08-31). **Reproducibility split:** the MAESTRO/Groove arms run from the deposited package plus the sha256-pinned public corpus downloads; the GiantMIDI arm of the null calibration reads a derived onset

series committed in a private sibling repository and therefore does **not** run from the deposit alone.
- This paper's own record: Zenodo DOI 10.5281/zenodo.22180435 (published 2026-08-31).
- Bouchaud, J.-P., M. Potters, and M. Meyer. 2000. "Apparent Multifractality in Financial Time Series." *European Physical Journal B* 13 (3): 595-599. (`bouchaud2000`)
- Bryce, R. M., and K. B. Sprague. 2012. "Revisiting Detrended Fluctuation Analysis." *Scientific Reports* 2: 315. (`bryce2012`)
- Calvet, L. E., and A. J. Fisher. 2001. "Forecasting Multifractal Volatility." *Journal of Econometrics* 105 (1): 27-58. (`calvet2001`)
- Calvet, L. E., and A. J. Fisher. 2004. "How to Forecast Long-Run Volatility: Regime Switching and the Estimation of Multifractal Processes." *Journal of Financial Econometrics* 2 (1): 49-83. (`calvet2004`)
- Calvet, L. E., and A. J. Fisher. 2008. *Multifractal Volatility: Theory, Forecasting, and Pricing.* Academic Press. (`calvet2008`)
- Cox, D. R. 1955. "Some Statistical Methods Connected with Series of Events." *Journal of the Royal Statistical Society, Series B* 17 (2): 129-164. (`cox1955`)
- Gillick, J., A. Roberts, J. Engel, D. Eck, and D. Bamman. 2019. "Learning to Groove with Inverse Sequence Transformations." *ICML*. (Groove MIDI Dataset v1.0.0; CC BY 4.0.) (`gillick2019`)
- Hawthorne, C., et al. 2019. "Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset." *ICLR*. (`hawthorne2019`)
- Hennig, H., et al. 2011. "The Nature and Perception of Fluctuations in Human Musical Rhythms." *PLoS ONE* 6 (10): e26457. (`hennig2011`)
- Hu, K., P. Ch. Ivanov, Z. Chen, P. Carpena, and H. E. Stanley. 2001. "Effect of Trends on Detrended Fluctuation Analysis." *Physical Review E* 64: 011114. (`hu2001`)
- Kantelhardt, J. W., et al. 2002. "Multifractal Detrended Fluctuation Analysis of Nonstationary Time Series." *Physica A* 316 (1-4): 87-114. (`kantelhardt2002`)
- Lux, T. 2008. "The Markov-Switching Multifractal Model of Asset Returns: GMM Estimation and Linear Forecasting of Volatility." *Journal of Business & Economic Statistics* 26 (2): 194-210. (`lux2008`)
- Peng, C.-K., et al. 1994. "Mosaic Organization of DNA Nucleotides." *Physical Review E* 49 (2): 1685-1689. (`peng1994`)
- Räsänen, E., et al. 2015. "Fluctuations of Hi-Hat Timing and Dynamics in a Virtuoso Drum Track of a Popular Music Recording." *PLoS ONE* 10 (6): e0127902. (`rasanen2015`)
- Schreiber, T., and A. Schmitz. 1996. "Improved Surrogate Data for Nonlinearity Tests." *Physical Review Letters* 77 (4): 635-638. (`schreiber1996`)
- Theiler, J., et al. 1992. "Testing for Nonlinearity in Time Series: The Method of Surrogate Data." *Physica D* 58 (1-4): 77-94. (`theiler1992`)
- Zhang, K. K., Y. Sun, and H. Xiong. 2025. "Multifractal Comparison of Billboard and AI-Generated Music." *Proc. 33rd ACM International Conference on Multimedia (MM '25)*. doi:10.1145/3746027.3758168. (`zhang2025`)