NO SINGLE POINT SUFFICES: REPORTING REGULARIZATION LANDSCAPES IN FROZEN-REPRESENTATION EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Frozen-representation evaluation compares pretrained representations through probe performance, but the measured advantage depends on the evaluation protocol. We study signed comparison gaps with ridge probes on three music benchmarks and an exploratory NLP/vision extension, distinguishing shared-penalty paths from independent validation-based selection. In the primary MTG-Jamendo configuration the sampled ROC-AUC gap has range 0.04432 and endpoint contrast 0.04156 (nominal 95% artist-cluster-bootstrap CI [+0.00765, +0.07980], 10,000 resamples), remaining positive at all ten sampled coefficients; simultaneous bands admit a constant gap curve. Both inherit the declared grid's extent: trimming the upper bound to the two arms' sampled peak $\alpha = 6.0 \times 10^4$ reduces the endpoint contrast to +0.0060 and the range to 0.00877. Within $\alpha \leq 6.0 \times 10^4$ the gap is not distinguishable from a constant gap under the paired cluster bootstrap ($p = 0.81$); the movement the bootstrap resolves is the post-peak rise from 6.0×10^4 to 10^7. Restricting each arm to validation-near-optimal coefficients reveals a different reporting limitation: at $\varepsilon = 0.005$, two of 22 cross-domain comparisons — each with a singleton common-coefficient set of zero range — admit independently acceptable test gaps of both signs, while the other twenty remain sign-consistent; matching effective complexity leaves four of 22 comparisons sign-changing. These exploratory envelopes are descriptive, not uncertainty-adjusted ranking claims. Different evaluation questions are therefore not interchangeable: a narrow shared-coefficient summary can omit variation under independent choices.

CONTRIBUTIONS

At a shared penalty we report signed gap paths, finite-grid ranges, and paired endpoint contrasts, separating changes in measured margin from changes in sampled ordering. We then distinguish a coefficient validation-near-optimal for both arms from coefficients each arm accepts on its own: a post hoc reanalysis of saved scores supplies cases where a singleton common set masks a wider, sign-changing set of independently acceptable test comparisons. Finally we evaluate the accuracy-based selection rule against metric-specific empirical test-grid oracles under the recorded candidate-evaluation policy. The decomposition $\Delta_{\text{tuned}} - \Delta_{\text{oracle}} = s_{\text{B}} - s_{\text{A}}$ is an accounting identity; the contribution lies in the measured differences between these evaluation questions, not in claiming priority for regularization sweeps.

1 INTRODUCTION

Frozen-representation evaluation trains a probe on a fixed encoder and uses predictive performance as evidence about its features (Alain and Bengio, 2017; Hewitt and Liang, 2019; Belinkov, 2022). A comparison between two encoders concerns both the direction and magnitude of their performance difference, and both are conditional on task, split, preprocessing, probe family, metric, and regularization protocol. Prior work has established that probing outcomes depend on the probe as well as the representation: Hewitt and Liang (2019) separate information in a representation from what a

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

probe can memorize, Belinkov (2022, §4.2) notes that a representation ranking can flip under alternative measures, and RAPTOR (Gao et al., 2026) examines regularization within a validation-tuned probing procedure. These observations motivate the work below; they are not the contribution itself.

We ask how the signed gap $\Delta_{A,B}(\lambda) = P(A, \lambda) - P(B, \lambda)$ varies across a declared regularization grid with the protocol fixed, and what gap results when each arm instead selects its own coefficient on validation data. The title concerns the information omitted by a single reported point, not a universal failure of single-score evaluation: an ordering that persists across sampled coefficients and an advantage that changes in magnitude are compatible outcomes. Analysing paired differences directly makes the comparison and its uncertainty explicit, because two representations can respond differently to regularization and a stable ordering need not imply a stable advantage.

For each sweep we fix the data, split, frozen features, preprocessing, probe family, metric, and candidate grid, then vary only the penalty, summarizing the path by $\Delta(\lambda)$, its range V_λ, and the endpoint contrast `I`. A separate analysis selects a coefficient per arm on validation and compares the resulting gap with a finite-grid test-oracle reference; that gap need not lie on the shared-penalty path, because the two selected penalties can differ. Closed-form ridge probes on frozen audio embeddings are the primary testbed; additional music settings and an exploratory NLP/vision extension examine scope.

The primary MTG result shows the distinction: the sampled gap stays positive at all ten coefficients, yet its range is 0.04432 and its endpoint contrast +0.04156, an exploratory comparison whose pointwise intervals are not simultaneous (§5); trimming to the two arms' sampled peak $\alpha = 6.0 \times 10^4$ reduces these to 0.00877 and +0.0060, within which the gap is not distinguishable from a constant ($p = 0.81$). Sign changes follow in §6 and near-optimal restrictions in §7. Under independent validation selection, the absolute offset from the empirical finite-grid oracle gap has median 0.0064 and maximum 0.0457 across 22 cross-domain comparisons, with opposite signs in one (§8); two comparisons admit test gaps of both signs once the arms choose acceptable coefficients separately (§7), each with a singleton common-coefficient set, so a zero common-set range guarantees nothing about independent choices. Fixed-CLAP experiments with repeated grouped partitions and two track-level bootstraps give a second reading: the validation bootstrap shifts the selected coefficient between two adjacent grid values while the test-oracle shortfall takes only 0 and 0.0014095, illustrating that movement in the selected coefficient need not entail appreciable score loss.

The winner can persist at the sampled coefficients while the measured margin moves, and validation selection yields a third quantity; these are separate claims, evaluated one at a time here. We infer neither how often reporting is inadequate in the literature nor any measured downstream benefit from the recommendation.

2 Related Work

Probing outcomes reflect both the representation and the evaluation procedure. Hewitt and Liang (2019) show why probe accuracy alone does not separate encoded information from memorization, and Belinkov (2022, §4.2) notes that a ranking "may be flipped under alternative measures". We take this dependence as established and measure signed, task-specific gaps across a declared grid.

Pareto probing (Pimentel et al., 2020) already studies complexity–performance trade-offs and advocates reporting curves, and GULP (Boix-Adserà et al., 2022) defines ridge-parameterized, prediction-based representation distances. We extend these perspectives to task-specific signed gaps, separating shared-penalty paths from independently selected comparisons. In vision, Lee et al. (2023) examine how linear-evaluation configurations affect model comparisons, though the pretraining weight-decay settings discussed there are distinct from the probe-time penalty varied here. In audio, probe structure and token aggregation affect frozen-representation evaluation (Rauch et al., 2026), with HEAR and MARBLE as benchmark context (Turian et al., 2022; Yuan et al., 2023).

Reporting and model-selection work adds context: Dodge et al. (2019) advocate reporting search budgets, Tang et al. (2020) probe the estimator's limits, and Lucic et al. (2018) find that with a sufficient tuning budget most GAN variants reach comparable scores, so that a reported advantage can reflect the search rather than the method, Bouthillier et al. (2021) and Reimers and Gurevych (2017) study variability across experimental choices, and Gorman and Bedrick (2019) and Søgaard et al. (2021) examine split sensitivity — motivating the explicit split definitions and paired, group-

aware resampling used here. Selection-induced optimism (Cawley and Talbot, 2010; Varma and Simon, 2006) is why §8 treats finite-grid test maxima as descriptive, not deployable, references. Our near-optimal sets are narrower than Rashomon-set (Fisher et al., 2019; Marx et al., 2020), underspecification (D'Amour et al., 2022) and specification-curve (Simonsohn et al., 2020; Steegen et al., 2016) analyses; a one-dimensional sweep inherits none of their guarantees, and membership implies no test-set equivalence.

RAPTOR (Gao et al., 2026) selects ridge regularization on validation data within a probe-then-steer pipeline; Appendix D applies a RAPTOR-style procedure to our conditions as a procedural analysis.

3 Setup and Definitions

We write probe performance as $P = f(R, H, \lambda, n, F, M)$: `R` the representation, `H` the probe family, λ the declared regularization coordinate, `n` the fitting-sample count, `F` the split construction, and `M` the metric. Preprocessing and label definitions belong to the protocol as well. Within a sweep only the coefficient varies, and α denotes the coefficient entering the ridge normal equations. A shared λ means a shared declared coefficient, not a shared index in two candidate lists: In the cross-domain analysis the comparison coefficients are the stored solver values α (Appendix J).

Three evaluation questions recur: comparing arms at a shared nominal coefficient, after each arm selects its own coefficient on validation, and at matched effective complexity. The first two are analysed here; Appendix H covers the third. A shared nominal coefficient need not impose equal shrinkage when feature dimensions differ, so the shared path is a property of the specified representation–probe protocol, not an invariant measure of representation quality.

For the shared-penalty protocol define $\Delta(\lambda) = P_{\text{test}}(A, \lambda) - P_{\text{test}}(B, \lambda)$, the finite-grid range $V_\lambda = \max_{\lambda\in\Lambda} \Delta(\lambda) - \min_{\lambda\in\Lambda} \Delta(\lambda)$, and the endpoint contrast $I = \Delta(\lambda_{\max}) - \Delta(\lambda_{\min})$. The sign of `I` implies nothing about the signs at the endpoints, and neither quantity fixes behavior between or beyond the sampled values.

For independent selection let $S_{\mathrm{R}}(\lambda)$ be the validation score from a training-stage fit and $T^{\pi}_{\mathrm{R}}(\lambda)$ its test score under a candidate-evaluation policy π; in the 22-pair analysis S_{R} is accuracy for both reported metrics. After selection every candidate is refitted on training plus validation with the same solver coefficients, and both the selected score and each metric-specific oracle are read from that refitted family. Define $\hat{\lambda}_{\mathrm{R}} = \arg\max_{\lambda\in\Lambda_R} S_{\mathrm{R}}(\lambda)$, $Q^{\pi}_{\mathrm{R}} = \max_{\lambda\in\Lambda_R} T^{\pi}_{\mathrm{R}}(\lambda)$ and $s^{\pi}_{\mathrm{R}} = Q^{\pi}_{\mathrm{R}} - T^{\pi}_{\mathrm{R}}(\hat{\lambda}_{\mathrm{R}})$, so that $\Delta_{\text{tuned}} - \Delta_{\text{oracle}} = s_{\mathrm{B}} - s_{\mathrm{A}}$. Shortfalls are non-negative because the selected score belongs to the same candidate family as the oracle, but the oracle reads test data retrospectively and is not an admissible selection rule, so these shortfalls bound no future selection error.

Near-optimal sets are $\Lambda_{\mathrm{R}}(\varepsilon) = \{\lambda \in \Lambda_{\mathrm{R}} : S_{\mathrm{R}}(\lambda) \geq \max_\ell S_{\mathrm{R}}(\ell) - \varepsilon\}$ and the common set is $C_\varepsilon = \Lambda_{\mathrm{A}}(\varepsilon) \cap \Lambda_{\mathrm{B}}(\varepsilon)$: each arm is nonempty on a finite grid, but the intersection can be empty, and a singleton has zero range by construction. For independent choices the admissible set is $\Lambda_{\mathrm{A}}(\varepsilon) \times \Lambda_{\mathrm{B}}(\varepsilon)$, whose test gaps are enclosed by

$$J^{\pi}_{\varepsilon} = [\min_{a\in\Lambda_A(\varepsilon)} T^{\pi}_{\mathrm{A}}(a) - \max_{b\in\Lambda_B(\varepsilon)} T^{\pi}_{\mathrm{B}}(b), \max_{a\in\Lambda_A(\varepsilon)} T^{\pi}_{\mathrm{A}}(a) - \min_{b\in\Lambda_B(\varepsilon)} T^{\pi}_{\mathrm{B}}(b)].$$

This is the smallest closed interval containing the candidate-pair gaps; not every value inside it need be attainable, and it is not a confidence interval. Section 7 computes these envelopes retrospectively from saved curves at the original tolerances, without refitting. For centered, standardized features the effective degrees of freedom of the penalized linear component are $df(\alpha) = \sum_k \sigma^2_{\mathrm{k}}/(\sigma^2_{\mathrm{k}} + \alpha)$, excluding the unpenalized intercept; matching `df/d` is distinct from matching absolute `df`, and dividing the normal equations by a sample count to form α/n removes neither spectral nor preprocessing differences.

4 Protocol

The primary analysis uses MTG-Jamendo (Bogdanov et al., 2019) under official `split-0`: train 9,949 / validation 3,802 / test 4,231. DEAM (Aljanaki et al., 2017), with recorded counts 744 / 1,000 / 58, and Emotify (Aljanaki et al., 2016), five folds of 80 tracks over nine GEMS outputs (Zentner et al., 2008), supply additional music settings. These conventions are not pooled, and the 58-track

DEAM test set limits precision. Music analyses use artist-grouped procedures, which track-level splitting can overlook (Flexer, 2007; Gorman and Bedrick, 2019); the official split, constructed-validation variants, and repeated group partitions are reported separately.

Both encoders are used without fine-tuning: laion-clap (Wu et al., 2023), checkpoint `630k-audioset-best.pt`, and MERT-95M (RVQ-VAE) (Li et al., 2024), layer 8. Track embeddings use 10 s windows, a 5 s hop, and window-mean aggregation. The probe is ridge with per-column standardization and an unpenalized intercept; writing $Y_c = Y - \bar{y}$, its coefficient matrix is $W = (X_c^T X_c + \alpha I)^{-1} X_c^T Y_c$. This fixes the solver-coordinate convention (Appendix J). The primary grid holds ten logarithmically spaced solver coefficients over $\alpha \in [10^{-3}, 10^{7}]$; a separate 21-point grid samples the same range more densely without containing all ten original points, so their observed ranges do not establish convergence.

The primary path uses a paired cluster bootstrap over 4,231 test tracks in 327 artist-identified $split_{\text{group}}$ clusters; each replicate draws 327 clusters with replacement and applies the same rows to both arms and every coefficient, holding predictions fixed. This estimates evaluation-sample uncertainty conditional on the fitted probes and split. The primary calculation uses $B = 400$ and seed 20260926, with endpoint replicates $\Delta_b(\alpha_{\max}) - \Delta_b(\alpha_{\min})$; the alternative configuration uses a separate $B = 10{,}000$ run. The endpoint contrast and ten-point Bonferroni and max-type simultaneous bands are additionally recomputed for the primary configuration at $B = 10{,}000$ under the identical cluster construction (Appendix K): both counts give endpoint intervals excluding zero, and both counts' bands contain a constant gap curve. All reported intervals remain exploratory.

Figure 1 describes the primary configuration: native validation, the full training split, a ten-point grid, and twelve labels `sorted(TAG56)[:12]`. The alternative `dark` label configuration of Appendix A.1 replaces `ambiental` with `dark` and uses `B = 10,000` and seed 20260919; both of its reported bands belong to that configuration and neither is substituted for the primary one. The forest collection (§6), the matched MTG protocol comparison (Appendix C), and the multi-metric music counts (§7) are not pooled.

Label-permutation checks give a reference after disrupting the feature–label association. Two scope limits belong here: all forty comparisons are non-music, so the control carries no paired music evidence; three archived single-arm music checks are reported with the domain decomposition in Appendix A.2 and Figure 6. A separate archived NLP-AG batch of sixteen comparisons, fourteen of which fail, is not pooled with the forty because it uses a different probe specification and comparison family, its failures coming from the smallest archived effects. Implementation checks confirm three invariances: identical arms give $\Delta \equiv 0$, exchanging the arms negates Δ and leaves V_λ unchanged, and jointly reordering aligned samples, predictions and labels leaves point estimates unchanged.

5 Results I: Sensitivity Along the Path

Figure 1 reports the primary configuration and Table 1 lists these values.

Under the primary MTG protocol (CLAP versus MERT-95M, native validation, full-training-split fitting, `sorted(TAG56)[:12]`), the endpoint contrast is `I = +0.04156` (nominal 95% paired cluster-bootstrap CI `[+0.00765, +0.07980]` at `B = 10,000`; `[+0.00499, +0.07935]` at `B = 400`) and the full-grid sampled range is $V_\lambda = 0.04432$ (0.00877 up to the sampled peak). The gap runs from +0.03046 to +0.07478, stays positive at all ten coefficients, and is non-monotone; every nominal pointwise interval excludes zero. The endpoint contrast was chosen after inspecting the grid, so its interval ignores that data dependence; pointwise positivity is not a simultaneous claim over the sampled range.

Two further descriptive summaries are the largest deviation from the sampled median, 0.0396, and the largest adjacent change, 0.0194, from 6.0×10^4 to 7.7×10^5. Because V_λ ranges over a declared ten-point grid, it inherits that grid's extent: the two largest coefficients account for 0.03555 of the 0.04432 range. Trimming to the two arms' sampled peak gives $V_\lambda = 0.00877$, which the constant-gap null (Figure 2a) does not distinguish from a flat gap ($p = 0.81$); the movement it resolves is the post-peak rise, 0.03555, the range of the final three coefficients ($p = 0.003$). Appendix K reports each cut's trimmed value and a constant-gap null for every subset; Figure 2 shows both pieces.

Under that null the final two coefficients give $p = 0.005$ and the three from 4.6×10^3 to 7.7×10^5 give $p = 0.012$, against $p = 0.053$ for the full grid and $p = 0.8065$ for the plateau segment; over a common coverage window the nominal and matched-complexity paths agree (0.00793 against 0.00864, Appendix H).

Within the segment $\alpha \leq 4.6 \times 10^3$ the gap varies by 0.0074, one sixth of V_λ, while the two arms' absolute test ROC-AUC moves by approximately 0.056 and 0.059. A flat segment does not show that the arms are unchanged, and one coefficient does not license extrapolation across the sampled range; both arms peak at $\alpha = 6.0 \times 10^4$, outside that segment. At the top of the declared grid the arms have $df = 0.50$ and 0.74 (Table 1) — under one effective degree of freedom — while their ROC-AUC still reaches 98.7% and 94.0% of their peaks; the post-peak movement is therefore measured where the probe has almost no fitting freedom, MERT-95M falling 0.0457 from its peak, CLAP 0.0101.

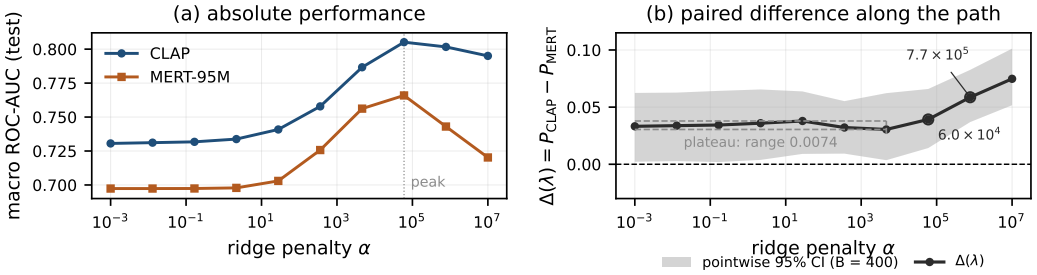

Figure 1: Primary MTG configuration containing ambiental under native validation, fitted on the full training split. (a) Absolute test ROC-AUC curves; both sampled maxima occur at $\alpha = 6.0 \times 10^4$. (b) The paired gap with nominal pointwise 95% intervals from 400 artist-cluster resamples. The bracket marks the value range for $\alpha \leq 4.6 \times 10^3$. Open circles highlight two intermediate coefficients above it; the final coefficient, 10^7, also exceeds it. These comparisons are descriptive; no simultaneous band is reported.

Table 1: Primary MTG pointwise summaries. CLAP versus MERT-95M on the twelve `sorted(TAG56)[:12]` labels at native validation, fitted on the full recorded training split. Intervals are nominal pointwise percentile intervals from 400 paired test-cluster resamples and are not simultaneous; the endpoint contrast uses only the first and last rows.

solver α	primary gap Δ	nominal 95% pointwise interval	df (CLAP / MERT-95M)
0.001	+0.033214	[+0.002192, +0.062435]	509 / 767
0.012915497	+0.033788	[+0.002753, +0.062720]	504 / 767
0.16681005	+0.034345	[+0.001937, +0.064233]	491 / 767
2.1544347	+0.035931	[+0.003807, +0.065468]	457 / 761
27.825594	+0.037866	[+0.009066, +0.063748]	375 / 704
359.38137	+0.032150	[+0.009300, +0.055245]	230 / 469
4641.5888	+0.030455	[+0.003559, +0.062232]	98.3 / 192
59948.425	+0.039223	[+0.014181, +0.065888]	30.3 / 49.8
774263.68	+0.058631	[+0.036899, +0.082600]	5.26 / 7.70
10000000	+0.074776	[+0.051668, +0.101188]	0.50 / 0.74

The alternative twelve-label configuration replaces `ambiental` with `dark`; its contrast is $I = V_\lambda = 0.04651$ with nominal 95% interval [0.01573, 0.07061] (Appendix A.1).

Table 2 and Figure 11 compare descriptive magnitudes. For the primary bootstrap, which resamples 327 artist clusters containing 4,231 test tracks, dividing percentile-interval half-widths by 1.96 gives normal-approximation scales of 0.019 for the endpoint contrast and 0.012–0.016 pointwise. These are interval-derived proxies, not directly estimated bootstrap standard deviations. A separate archived 56-label diagnostic gives an approximate $K = 5$ fold-SD proxy of 0.013–0.015; its reconstruction substitutes a full-test gap for the fold mean (supplementary `metadata/s15_fold_sd_derivation.md`). The across-label SD, 0.042–0.053, describes

heterogeneity. None of these scale comparisons tests whether sampling variation could produce the observed path movement, and the fold diagnostic is not independent significance evidence.

Table 2: Primary-grid statistics and descriptive reference scales, in ROC-AUC units. Bootstrap scales are percentile-interval half-widths divided by 1.96; the archived fold SD is a reconstructed proxy. Ratios are descriptive and provide no significance test.

quantity	value	descriptive ratio
shared-path range V_λ	0.04432	—
endpoint contrast `I`	0.04156	—
max deviation from the sampled median (grid-defined)	0.0396	—
max adjacent step (grid-defined)	0.0194	—
V_λ without the final coefficient (B1)	0.02818	—
V_λ without the final two coefficients (B2)	0.00877	—
archived fold-SD proxy ($K = 5$)	0.013–0.015	—

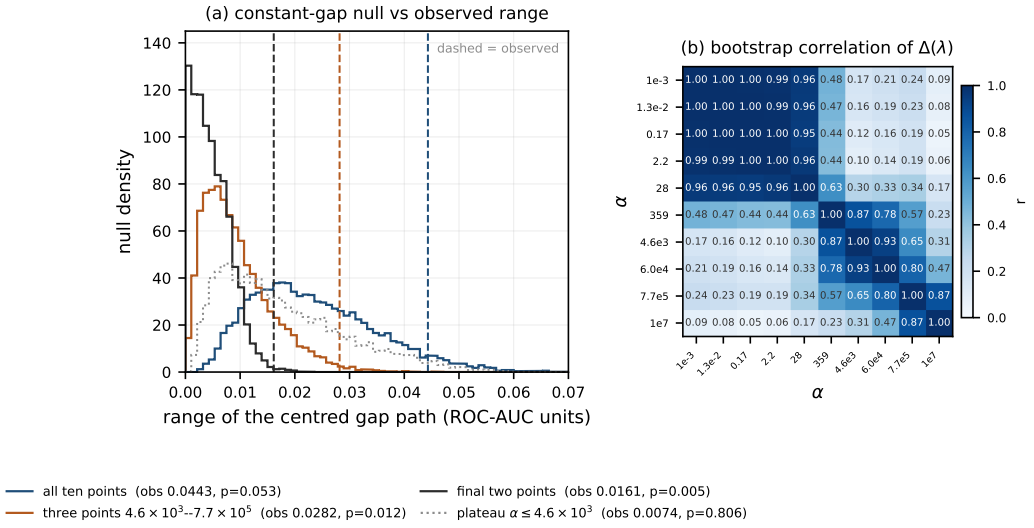

Figure 2: Primary-configuration null calibration at $B = 10{,}000$. (a) Constant-gap null distributions of the centred path range for four coefficient subsets; dashed lines mark the observed ranges. (b) Bootstrap correlation matrix of $\Delta(\lambda)$ across the ten coefficients. Correlation decays with α separation (median $r = 0.93$ for adjacent coefficients, one grid step apart, and 0.17 beyond five decades), which is why the eight-coefficient range is small while the endpoint pair is nearly uncorrelated ($r = 0.093$).

6 Results II: Descriptive Sign Changes Across Conditions

Sampled gap estimates change sign across the reported conditions; these counts are separate from inferential claims about ranking reversals.

Figure 3 summarizes twelve archived conditions: eleven MTG macro ROC-AUC comparisons under a constructed validation split, plus a DEAM R^2 comparison. Each row reports the endpoint contrast with its nominal paired-bootstrap interval (`B = 400`); the sign-change marker records sampled point-estimate sign changes, and Table 6 in Appendix A lists the twelve rows. The DEAM row has 58 evaluation tracks, and its R^2 contrast is not on a common effect scale with the MTG AUC contrasts;

Evaluation sets differ across conditions: the main backbone comparisons use 4,231 test tracks, the layer and window comparisons a 357-track subset, and the MERT-330M comparison 2,810 tracks with embeddings. Only the constructed-validation CLAP–MERT-95M interval excludes zero; the other eleven contain zero, and an interval containing zero is not evidence of no effect. Six of the

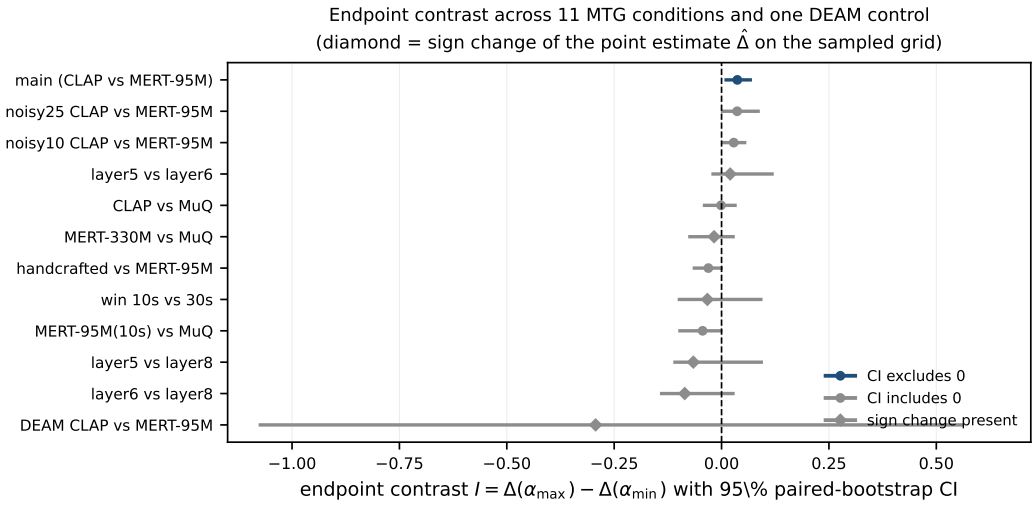

Figure 3: Archived twelve-condition endpoint contrasts with nominal 95% paired-bootstrap intervals. A diamond denotes a descriptive sign change in the sampled gap estimates, not a statistically established ranking reversal. The MTG rows use the constructed-validation batch, not the primary protocol.

twelve conditions show a sign change in the sampled point estimates, ROC-AUC among them; all six evaluate 3,810 or fewer tracks (four on the 357-track subsets, one on 2,810, one on 58), while none of the six 4,231-track conditions changes sign. In four of the six every relevant pointwise interval contains zero, leaving the local signs unresolved; the reverse also occurs. Advantage at a point and change in advantage are different estimands, and none of this amounts to a supported reversal.

A paired rerun of the twelve matched MTG entries under native validation holds each test set fixed and omits the DEAM row; the sampled sign-change count holds in 11 of 12 entries and moves from 0 to 1 in the last, so the pattern survives the joint protocol change, though that change is not isolated causally because the fitting set moves with the validation construction.

The archived fold statistic also varies with the penalty (2.8–8.5 over $K = 5, 10, 20$); it reuses fitted predictions and is not independent evidence for the primary path.

7 Results III: Shared and Independently Acceptable Coefficients

A full grid can contain coefficients far from either arm's validation optimum, so we first restrict each shared path to C_ε, the intersection of the two near-optimal sets. An empty intersection has no defined range and a singleton has zero range by construction, but neither implies that the arms lack separately acceptable choices. Table 10 in Appendix A reports the resulting set sizes and the counting unit behind each one.

In the constructed batch all 30 multi-point condition–tolerance entries belong to MTG macro ROC-AUC, label-noise conditions included; the DEAM row is a singleton at every tolerance and contributes none. Across those 30 entries the median full-grid range is 0.04810 against a restricted median of 0.00951, a ratio of 5.06 between two medians over repeated entries, which is neither a condition-wise median nor an estimate from 30 independent experiments. Five of the twelve native MTG entries have empty intersections at $\varepsilon = 0.005$, the primary condition among them.

The 22-pair archive also supports a consistently refitted candidate-family analysis with acceptable coefficients defined by validation **accuracy** for both test metrics. The cross-domain account keeps tolerances apart and restricts each pair of range medians to the same multi-point subset (Table 3, Appendix A).

At $\varepsilon = 0.005$ the median full-grid range over the multi-point subset is 0.44164 against a common-set median of 0.00538; at 0.01, 0.32610 against 0.00400; at 0.02, 0.26225 against 0.00915 (empty/single/multi counts 5/12/5, 5/8/9, 4/2/16). Restriction shrinks the observed range in the displayed subsets without yielding one cross-domain factor. For every nonempty set $0 \le V_{\text{common}} \le V_{\text{full}}$ follows from ranging over a subset, so what matters empirically is the magnitude and where it falls; membership on validation gives no test equivalence and the counts estimate no shared empty-set probability.

The shared intersection omits pairs in which the two arms use different acceptable coefficients, so we also reanalyse the saved validation and refitted test curves for all 22 cross-domain comparisons at the same three tolerances. This adds no model fitting; it summarises stored results and is neither pre-registered nor an uncertainty interval. Validation here holds 79–400 examples, so ε is below one validation standard error (Appendix D); the both-sign outcome nevertheless recurs at all three tolerances. The accuracy-gap envelopes have median width 0.00815, 0.01153 and 0.03200 at the three tolerances and carry both signs in 2, 2 and 3 of the 22 cases. Two AG News pretraining comparisons show a singleton common intersection coexisting with independently acceptable test gaps of both signs, so a common-set range of exactly zero does not certify a stable direction (Appendix A.3); at $\varepsilon = 0.005$ the other twenty envelopes have a strictly consistent sign.

Let w_R be the range of T_R^π over $\Lambda_R(\varepsilon)$; subtracting the endpoints in Section 3 gives $\mathrm{width}(J_\varepsilon^\pi) = w_A + w_B$. A shared-coefficient gap can nevertheless remain constant when both scores move together. Suppose both arms have the same acceptable set and $T_A^\pi(\lambda) = T_B^\pi(\lambda) + c$ throughout it. The common-set gap is always c, whereas the independent-choice envelope is $[c - w, c + w]$, where w is either arm's score range. It contains gaps of both signs whenever $w > |c|$. Thus even a multi-point common set with zero gap range need not imply stability under independent choices. The archived AG News cases establish its occurrence under the examined validation tolerances.

Because each arm's validation maximizer belongs to its own near-optimal set, $\Delta_{\text{tuned}} \in J_\varepsilon^\pi$ for every $\varepsilon \ge 0$ under a fixed candidate-evaluation policy; the envelope asks whether that direction persists across the admitted alternatives.

8 Results IV: Consequences of Validation-Based Selection

The second estimand is $\Delta_{\text{tuned}} = T_{\text{A}}^\pi(\hat{\lambda}_{\text{A}}) - T_{\text{B}}^\pi(\hat{\lambda}_{\text{B}})$: a coefficient selected separately for each arm on validation data, with test performance under the policy π of §3. Its difference from Δ_{oracle} is an empirical selection offset; the oracle serves only retrospectively.

The 22 pairs comprise nine UD English-EWT part-of-speech, four AG News and nine CIFAR-10 comparisons; Appendix D identifies both arms and sample pools. Every arm selects by validation accuracy, breaking ties toward the smaller coefficient, and all candidates are refitted on training plus validation. Selected-coefficient ratios have median 10 and maximum 10^6, exceeding 10^3 in six pairs; these describe differences between representations, not resampling instability.

For accuracy the median absolute gap offset is 0.0064 and its observed maximum 0.0457; for macro-F1 they are 0.0085 and 0.0431. Because $\Delta_{\text{tuned}} - \Delta_{\text{oracle}} = s_{\text{B}} - s_{\text{A}}$ with both shortfalls non-negative, a paired offset cannot exceed the larger single-arm shortfall, and the accuracy maximum coincides with the largest single-arm shortfall over 44 arms, 0.0457, so that pair's other arm has zero shortfall: an internal consistency check of the identity rather than an additional finding. Both metrics have opposite selected and oracle gap signs in the same condition, $AG_{\text{pretrain}}_n1000$ — descriptive signs, not statistically established reversals. Relative offsets, computed within each condition as $|\Delta_{\text{tuned}} - \Delta_{\text{oracle}}|/|\Delta_{\text{oracle}}|$, have medians 2.6% and 3.8% and maxima 312% and 291%: percentages of the comparison gap, not losses in predictive accuracy, and small denominators produce large values. An empirical test oracle does not turn the absolute paired offset into a conservative bound on future performance, because the two arms' maximization biases can affect their difference in either direction. Table 5 in Appendix A tabulates both metrics, and Figure 4 in that appendix displays the absolute offset against the reference gap on log–log axes.

A fixed-CLAP MTG extension draws eight 60/20/20 grouped partitions at requested training caps of 9,949 and 2,000. Shortfalls are measured against the observed grid maximum, so they are not population bounds and say nothing about variability of a two-representation comparison; they are zero in 6 of 8 and 4 of 8 partitions, with medians 0.000000 and 0.000128 and maxima 0.001585 and

0.004555. A separate perturbation design on the official split (`B = 500`, seed 424242) resamples test rows with the selection fixed, or validation rows with the test oracle fixed: median shortfalls 0.001415 and 0.001410, with the selected location moving while the score barely does — in the validation bootstrap 386 of 500 replicates select 3×10^5 and 114 select 10^5, whose fixed-test shortfalls are 0.0014095 and zero. Both designs resample tracks, not artist clusters (Tables 17 and 18).

These observations do not show that routine validation selection is generally unreliable: most near-optimal envelopes retain one sign and the fixed-CLAP shortfalls are small. Split maxima, track-bootstrap quantiles and paired offsets answer different questions; none bounds population risk.

For independently tuned comparisons, declare the coefficient convention, candidate grid, validation metric, tie-breaking rule and refitting policy. Specify the near-optimal tolerance without consulting test outcomes, and retain candidate-level validation and test scores. Alongside the selected gap, report each acceptable set's size and the endpoints of J^{π}_{ε}; when reporting a common-set range, also give the intersection size so that singleton zeros are identifiable. With candidate scores available, the envelope needs only four extrema and no added fitting. These summaries distinguish stability of the selected comparison from stability across acceptable alternatives. Population-level ordering claims additionally require paired uncertainty appropriate to the stated comparison; the envelope itself describes the recorded candidates.

9 Limitations

Its nominal interval carries a data-dependent endpoint choice and no multiplicity correction; the two resampling counts give lower bounds of 0.0050 and 0.0077, and the ten-point Bonferroni and max-type bands at $B = 10{,}000$ both contain a constant gap curve, so the contrast remains exploratory and is not the paper's inferential pillar. The archived `dark`-configuration bands show the same pattern. Sampled or candidate-pair sign differences do not establish statistically supported reversals, and non-nesting ten- and 21-point grids do not prove convergence at unsampled coefficients.

The candidate-pair envelopes are an exploratory reanalysis of stored scores: they carry no paired sampling intervals and leave the test set inadmissible as a selection source, so a sign-consistent finite envelope is not a stability guarantee. Shared nominal penalties, normalized-`df` matching, and independently selected coefficients implement different comparisons; the auxiliary matched-`df` analysis is confined to recorded common reachable targets, with a primary-configuration recomputation in Appendix H, and neither convention supplies a representation-invariant normalization.

The primary result concerns one corpus, one encoder pair, and twelve specified labels. Full-label and other music configurations broaden the observations without establishing population-level generality, and reused data, folds, grids, and condition–tolerance entries are not independent studies. The fixed-CLAP split analysis records caps, and S-42c resampling does not preserve artist clusters. The single-arm logistic check is not a paired-gap replication, and frozen ridge results do not establish behavior under fine-tuning or other audio time scales. The study does not measure the downstream decision benefit or computational cost of sensitivity reporting.

10 Conclusion

A representation comparison can retain the same sampled winner while its measured advantage changes. In the primary MTG configuration the ROC-AUC gap stays positive at all ten sampled coefficients, but its range is 0.04432 and exploratory endpoint contrast 0.04156, falling to 0.00877 and +0.0060 once the upper bound is trimmed to the two arms' sampled peak $\alpha = 6.0 \times 10^4$. These results describe declared grids and protocols, not a universal failure of frozen-representation evaluation.

A shared near-optimal coefficient and two independently near-optimal ones are different restrictions: in two AG News comparisons at $\varepsilon = 0.005$ a singleton common intersection has zero range, yet independently acceptable pairs give test gaps of both signs; the 22-pair accuracy-selected evaluation has median absolute oracle-relative offset 0.0064 and maximum 0.0457. Neither constitutes an uncertainty-adjusted ranking claim.

A selected probe score remains useful for its specified protocol.

AI USE STATEMENT

We used generative AI tools as described below. All scientific claims, experimental decisions, and numerical results were checked by the author against the archived artifacts, and the author remains responsible for them.

Formatting. Generative AI tools were used for formatting the manuscript to the ICLR 2027 LaTeX style requirements: section heading levels (the three closing statements are unnumbered subsections), table column alignment and column-width allocation, and reference formatting. All such changes were reviewed by the author against the official ICLR style guidelines.

Retrieval and discovery. Generative AI tools were used for retrieval and discovery of related literature, specifically for suggesting relevant papers in frozen-representation evaluation, including complexity–performance reporting and ridge-parameterized representation distances. All AI-suggested sources were independently verified by the author through manual literature searches on Google Scholar, Semantic Scholar, and DBLP. Each of the 35 references was confirmed to exist and to support the claim attributed to it; the bibliography has no uncited or dangling entries.

Design of methods and experiments. Generative AI tools were used to suggest experimental configurations and parameter ranges: the candidate regularization grids, the paired cluster-bootstrap resampling design for the primary sweep, the tolerances and candidate-evaluation policies for the near-optimal and selection analyses, and the grouped-partition and perturbation designs for the fixed-encoder extension. AI tools also proposed candidate hypotheses, which were treated as candidates to be falsified. All such designs were reviewed by the author, and final decisions were made from pilot runs, predeclared criteria, and available compute.

Implementation. Generative AI tools were used to implement methods, including experiment, analysis, and verification scripts for loading from the archived JSON and CSV artifacts, the closed-form ridge probe and its effective-degrees-of-freedom diagnostic, the paired artist-cluster bootstrap, the label-permutation control, and the recalculation of the reported summaries and figure data. All AI-generated code was tested by the author and entered the analysis only after passing an internal suite of 178 assertions. An independently implemented reference calculation agrees with scikit-learn's `linear_model.Ridge` to 3.56×10^{-16} relative error in predictions and 2.04×10^{-16} in the objective (predeclared tolerances 10^{-8} and 10^{-10}), and a deliberately corrupted replicate fails 36 of them. Replay reproduces the archived scores bitwise in the recorded environment.

Interpretation of results. Generative AI tools were used for interpreting results, specifically for suggesting explanations of observed phenomena: the flat low-regularization segment of the primary gap path despite larger movement in each arm's absolute performance; the dependence of the sampled peak coefficient on the reported metric; and the small discrepancy between the two near-identical label configurations. All AI-suggested interpretations were cross-checked against the experimental logs and archived artifacts; only those consistent with the raw data were included. Two auxiliary analyses and several proposed scaling relationships were withdrawn on that basis and are documented in Appendix F. AI was not used to generate results or any reported number.

Writing. Generative AI tools were used to aid and polish the writing — grammar, wording, clarity, and formatting — and to draft portions of the manuscript, including the abstract. Those drafted portions were revised and rewritten by the author, and no viewpoint or claim originated with AI.

ETHICS STATEMENT

This study analyzes existing public music, language, and vision benchmarks using frozen representations. It involves no newly recruited participants and no newly collected data. The primary MTG analysis uses the stated official split; constructed-validation variants and additional partitions are described separately. Results are presented within the limitations of the original datasets and the specified evaluation protocols.

REPRODUCIBILITY STATEMENT

The supplement provides per-candidate JSON results, summary CSV files, analysis scripts and configuration fingerprints. The evidence review recomputes summaries without new model training;

the product-set envelopes in §7 remain post hoc. Cross-domain identities appear in Appendix D; the supplement records grid conventions and fitting scripts. Primary `ambiental` results and alternative `dark` results remain separate. The fold-SD proxy is reconstructed as documented in `metadata/s15_fold_sd_derivation.md`; it is not an archived direct estimate. Access limits for raw data, embeddings and prediction caches are stated in `data/prediction_or_embedding_access.md`.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In International Conference on Learning Representations (ICLR), Workshop Track, 2017. arXiv:1610.01644 (first circulated 2016).

Anna Aljanaki, Frans Wiering, and Remco C. Veltkamp. Studying emotion induced by music through a crowdsourcing game. Information Processing & Management, 52(1):115–128, 2016. doi: 10.1016/j.ipm.2015.03.004.

Anna Aljanaki, Yi-Hsuan Yang, and Mohammad Soleymani. Developing a benchmark for emotional analysis of music. PLOS ONE, 12(3):e0173392, 2017. doi: 10.1371/journal.pone.0173392.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. Computational Linguistics, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422. arXiv:2102.12452v4.

Yoshua Bengio and Yves Grandvalet. No unbiased estimator of the variance of K-fold cross-validation. Journal of Machine Learning Research, 5:1089–1105, 2004.

Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. The MTG-Jamendo dataset for automatic music tagging. In Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML), 2019. URL: http://hdl.handle.net/10230/42015.

Enric Boix-Adserà, Hannah Lawrence, George Stepaniants, and Philippe Rigollet. GULP: a prediction-based metric between representations. In Advances in Neural Information Processing Systems (NeurIPS), volume 35, pp. 7115–7127, 2022. arXiv:2210.06545.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gael Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In Proceedings of Machine Learning and Systems (MLSys), volume 3, pp. 747–769, 2021.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research, 11:2079–2107, 2010.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. Journal of Machine Learning Research, 23(226):1–61, 2022.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 2185–2194, 2019. doi: 10.18653/v1/D19-1224.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. Journal of Machine Learning Research, 20(177):1–81, 2019.

Arthur Flexer. A closer look on artist filters for musical genre classification. In Proceedings of the 8th International Society for Music Information Retrieval Conference (ISMIR), 2007.

Ziqi Gao, Yaotian Zhu, Qingcheng Zeng, Xu Zhao, Ziqing Wang, Feng Ruan, and Kaize Ding. RAPTOR: Ridge-adaptive logistic probes. arXiv preprint arXiv:2602.00158, 2026. Version 2.

Kyle Gorman and Steven Bedrick. We need to talk about standard splits. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), 2019.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275. arXiv:1909.03368.

Torsten Hothorn, Frank Bretz, and Peter Westfall. Simultaneous inference in general parametric models. Biometrical Journal, 50(3):346–363, 2008. doi: 10.1002/bimj.200810425.

Jae-Hun Lee, Doyoung Yoon, ByeongMoon Ji, Kyungyul Kim, and Sangheum Hwang. Rethinking evaluation protocols of visual representations learned via self-supervised learning. arXiv preprint, 2023. arXiv:2304.03456.

Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghao Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, Norbert Gyenge, Roger Dannenberg, Ruibo Liu, Wenhu Chen, Gus Xia, Yemin Shi, Wenhao Huang, Zili Wang, Yike Guo, and Jie Fu. MERT: Acoustic music understanding model with large-scale self-supervised training. In International Conference on Learning Representations (ICLR), 2024. arXiv:2306.00107 (first circulated 2023).

Lawrence I-Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. Biometrics, 45(1):255–268, 1989. doi: 10.2307/2532051.

Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are GANs created equal? A large-scale study. In Advances in Neural Information Processing Systems (NeurIPS), volume 31, 2018. arXiv:1711.10337.

Charles Marx, Flavio du Pin Calmon, and Berk Ustun. Predictive multiplicity in classification. In Proceedings of the 37th International Conference on Machine Learning (ICML), volume 119 of PMLR, pp. 6765–6774, 2020.

Tiago Pimentel, Naomi Saphra, Adina Williams, and Ryan Cotterell. Pareto probing: Trading off accuracy for complexity. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 3138–3153, 2020. doi: 10.18653/v1/2020.emnlp-main.254. arXiv:2010.02180.

Lukas Rauch, René Heinrich, Houtan Ghaffari, Lukas Miklautz, Ilyass Moummad, Bernhard Sick, and Christoph Scholz. Unmute the patch tokens: Rethinking probing in multi-label audio classification. In International Conference on Learning Representations (ICLR), 2026. arXiv:2509.24901.

Nils Reimers and Iryna Gurevych. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2017.

Uri Simonsohn, Joseph P. Simmons, and Leif D. Nelson. Specification curve analysis. Nature Human Behaviour, 4(11):1208–1214, 2020. doi: 10.1038/s41562-020-0912-z.

Anders Søgaard, Miryam de Lhoneux, Yova Kementchedjhieva, and Johannes Bjerva. We need to talk about random splits. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL), 2021.

Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. Increasing transparency through a multiverse analysis. Perspectives on Psychological Science, 11(5):702–712, 2016.

Raphael Tang, Jaejun Lee, Ji Xin, Xinyu Liu, Yaoliang Yu, and Jimmy Lin. Showing your work doesn't always work. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 2766–2772, 2020. doi: 10.18653/v1/2020.acl-main.246. arXiv:2004.13705.

Joseph Turian, Jordie Shier, Humair Raj Khan, Bhiksha Raj, Björn W. Schuller, Christian J. Steinmetz, Colin Malloy, George Tzanetakis, Gissel Velarde, Kirk McNally, Max Henry, Nicolas Pinto, Camille Noufi, Christian Clough, Dorien Herremans, Eduardo Fonseca, Jesse Engel, Justin Salamon, Philippe Esling, Pranay Manocha, Shinji Watanabe, Zeyu Jin, and Yonatan Bisk. HEAR: Holistic evaluation of audio representations. In Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track, volume 176 of Proceedings of Machine Learning Research, pp. 125–145, 2022.

Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics, 7(91), 2006.

Gaël Varoquaux. Cross-validation failure: Small sample sizes lead to large error bars. NeuroImage, 180:68–77, 2018.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10095969.

Ruibin Yuan, Yinghao Ma, Yizhi Li, Ge Zhang, Xingran Chen, Hanzhi Yin, Le Zhuo, Yiqi Liu, Jiawen Huang, Zeyue Tian, Binyue Deng, Ningzhi Wang, Chenghua Lin, Emmanouil Benetos, Anton Ragni, Norbert Gyenge, Roger Dannenberg, Wenhu Chen, Gus Xia, Wei Xue, Si Liu, Shi Wang, Ruibo Liu, Yike Guo, and Jie Fu. MARBLE: Music audio representation benchmark for universal evaluation. In Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track, volume 36, 2023. arXiv:2306.10548.

Marcel Zentner, Didier Grandjean, and Klaus R. Scherer. Emotions evoked by the sound of music: characterization, classification, and measurement. Emotion, 8(4):494–521, 2008. doi: 10.1037/1528-3542.8.4.494.

A Cross-Condition and Cross-Domain Summary Statistics

Cross-domain tables referenced from §7 and §8. Table 3 gives the cross-domain common sets and paired test-accuracy ranges, Table 4 the accuracy-gap envelope widths by tolerance, and Table 5 the tuned-versus-oracle selection offsets; Figure 4 displays those offsets against the reference gap on log–log axes. All four use the same saved 22-pair archive and add no fitting.

Table 3: Cross-domain common sets and paired test-accuracy ranges. Empty / single / multi counts sum to 22 in every row. The last two columns use only that row's multi-point comparisons.

validation-accuracy tolerance ε	empty / single / multi	full-grid range, median on multi-point subset	common-set range, median on the same subset
0.005	5 / 12 / 5	0.44164	0.00538
0.01	5 / 8 / 9	0.32610	0.00400
0.02	4 / 2 / 16	0.26225	0.00915

Table 6 lists the twelve archived conditions of Figure 3 with their nominal paired-bootstrap intervals; it is the tabular form of that figure, and the interval column is nominal and unadjusted.

Table 4: Envelope widths by validation-accuracy tolerance. All 22 cross-domain comparisons enter every row; the last column counts envelopes whose endpoints carry both signs.

validation-accuracy tolerance ε	paired comparisons	median accuracy-envelope width	maximum accuracy-envelope width	envelopes with strictly negative and positive endpoints
0.005	22	0.00815	0.08440	2 / 22
0.01	22	0.01153	0.08440	2 / 22
0.02	22	0.03200	0.09330	3 / 22

Table 5: Tuned-versus-oracle selection offsets by metric. Percentage columns are percentages of the comparison gap, not of predictive accuracy.

metric	paired comparisons	median absolute offset	observed maximum absolute offset	opposite gap signs (tuned vs oracle)	median reported relative offset	maximum reported relative offset
accuracy	22	0.0064	0.0457	1 of 22	2.6%	312%
macro-F1	22	0.0085	0.0431	1 of 22	3.8%	291%

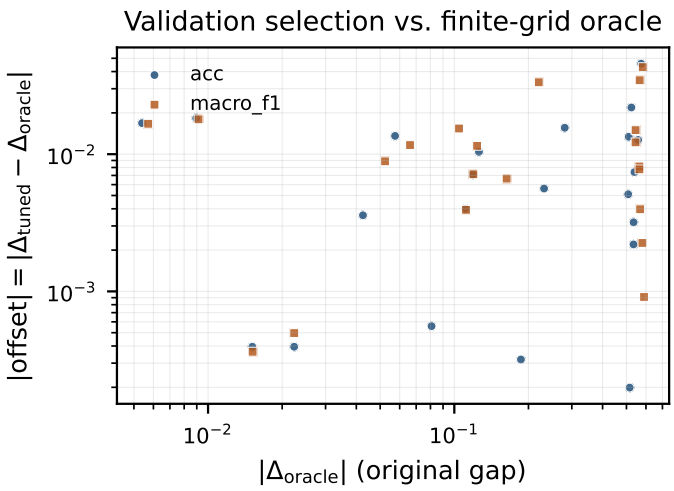

Figure 4: Absolute selection offset versus reference-gap magnitude, displayed by metric on log–log axes. Relative offsets can be large for near-tied comparisons. Opposite signs of the tuned and oracle gaps are descriptive and do not establish statistically supported reversals.

A.1 Archived grid-wide swings across music conditions

Table 7 retains an earlier archived set of nine music conditions, distinct from the §6 collection; its full-split twelve-label row uses the `dark` configuration. The 56-label S9 macro-AUC row uses the thirteen-point S9 grid from 10^{-3} to 10^{6}, while the macro-AUC, micro-AUC, and PR-AUC rows use the S-37 coefficients $\{1, 300, 10^5\}$. S9 and the primary twelve-label list are independently defined rather than nested, so these rows are neither one grid nor interchangeable replications. DEAM R^2 is the only non-rank row here.

Within this archived table, the eight rank-based rows have median $V_\lambda = 0.0192$ and range $[0.0050, 0.1031]$, with 0 of 8 showing a sign change. The single non-rank row has $V_\lambda = 0.2781$ and a sign change. These counts apply only to these rows. They do not supersede the later ROC-AUC sign-changing conditions in §6 or establish that rank-based metrics prevent sign changes.

Table 6: Twelve archived conditions with nominal 95% paired-bootstrap intervals. The interval column is nominal and unadjusted; sign-change counts are descriptive.

condition	n_{test}	$\hat{V}_\lambda$	I	nominal 95% CI for I	interval excludes 0	sampled sign-change count
CLAP vs MERT-95M (constructed val)	4,231	0.04874	+0.03673	[+0.00669, +0.07075]	yes	0
layer5 vs layer8	357	0.09028	−0.06571	[−0.11233, +0.09639]	no	3
layer6 vs layer8	357	0.09456	−0.08576	[−0.14340, +0.03039]	no	1
layer5 vs layer6	357	0.03017	+0.02005	[−0.02375, +0.12151]	no	1
MERT-330M vs MuQ	2,810	0.04662	−0.01748	[−0.07767, +0.03073]	no	2
window 10 s vs 30 s	357	0.04687	−0.03306	[−0.10214, +0.09535]	no	1
MERT-95M (10 s) vs MuQ	4,231	0.04410	−0.04410	[−0.10084, +0.00214]	no	0
CLAP vs MuQ	4,231	0.01285	−0.00129	[−0.04353, +0.03533]	no	0
handcrafted vs MERT-95M	4,231	0.07440	−0.03073	[−0.06722, +0.00330]	no	0
noisy10 CLAP vs MERT-95M	4,231	0.04795	+0.02835	[−0.00068, +0.05770]	no	0
noisy25 CLAP vs MERT-95M	4,231	0.04889	+0.03634	[−0.00163, +0.08910]	no	0
DEAM CLAP vs MERT-95M (R^2)	58	0.71857	−0.29322	[−1.07780, +0.56493]	no	1

Table 7: Archived grid-wide swings across nine music conditions. The 56-label S9 macro-AUC row uses the thirteen-point S9 grid from 10^{-3} to 10^{6}, while the macro-AUC, micro-AUC and PR-AUC rows use the S-37 coefficients $\{1, 300, 10^5\}$. These rows are neither one grid nor interchangeable replications, and the DEAM R^2 row is the only non-rank entry.

dataset	metric	Δ range	V_λ	sign change
MTG 12-dim, n = 9,949	macro AUC (rank)	[+0.0265, +0.0730]	0.0465	no
MTG 12-dim, n = 2,000	macro AUC (rank)	[+0.0410, +0.0659]	0.0249	no
MTG 12-dim, n = 1,000	macro AUC (rank)	[+0.0368, +0.0751]	0.0383	no
MTG 12-dim, n = 400	macro AUC (rank)	[+0.0157, +0.1188]	0.1031	no
MTG 56-dim	macro AUC (rank)	[+0.0180, +0.0315]	0.0135	no
MTG 56-dim	PR-AUC (rank)	[+0.0131, +0.0220]	0.0089	no
MTG 56-dim	micro AUC (rank)	[+0.0122, +0.0187]	0.0065	no
MTG 56-dim	macro AUC (rank)	[+0.0202, +0.0252]	0.0050	no
DEAM	R^2 (non-rank)	[−0.0042, +0.2739]	0.2781	yes

A.2 Cross-domain extension and the label-permutation control

Table 9 summarizes the broader archived cross-domain path analysis. Its counts are condition–metric entries, not the 22 paired comparisons used for independent validation selection in §8, so the two aggregation units stay apart. `A` denotes the full recorded grid and `B1` the same grid without its final point; both are analysis variants, not the two representation arms. Metric-family grouping organizes the summaries without making different tasks or metrics interchangeable.

The archived family-level summaries give medians of 0.0097 (A) / 0.0074 (B1) over 16 rank-based entries and 0.2210 (A) / 0.0790 (B1) over 64 threshold-based entries. These broad aggregates are descriptive, not a common-scale estimate across metrics or independent replications. Within these archived collections, sampled sign-change counts are 0/8 for music rank-based rows, 0/16 for cross-domain rank-based entries, 1/1 for the music non-rank row, and 8/64 for cross-domain non-rank

Table 8: Label-subset sensitivity configuration, not the primary numerical result. This configuration contains `dark` in place of `ambiental` and is the configuration displayed in Figure 5. It is retained because §5 uses its label-subset difference as a controlled comparison against the primary result.

α	$\Delta(\lambda)$	pointwise 95% CI	excludes 0
10^{-3}	+0.026527	[+0.005238, +0.051997]	yes
1.2916×10^{-2}	+0.027025	[+0.005871, +0.052458]	yes
0.16681	+0.027911	[+0.006847, +0.053082]	yes
2.15443	+0.030354	[+0.009802, +0.055133]	yes
27.8256	+0.033298	[+0.013108, +0.056649]	yes
359.381	+0.030798	[+0.008460, +0.051120]	yes
4.6416×10^{3}	+0.026980	[+0.001679, +0.049316]	yes
5.9948×10^{4}	+0.034832	[+0.011307, +0.054935]	yes
7.7426×10^{5}	+0.056945	[+0.036178, +0.075601]	yes
10^{7}	+0.073035	[+0.051224, +0.091972]	yes

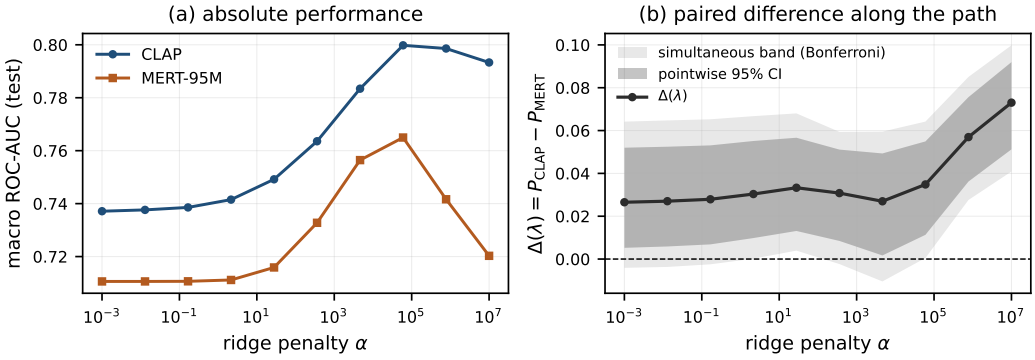

Figure 5: Alternative twelve-label MTG configuration containing dark, fitted on the full training split. (a) Absolute test ROC-AUC curves. (b) The paired gap, pointwise 95% intervals, and nominal Bonferroni simultaneous band from 10,000 test-artist-cluster resamples. These data have I = V_λ = 0.04651; they are not the ambiental primary configuration with I = 0.04156.

Table 9: Cross-domain grid-range medians by domain and metric family. The counts are condition–metric entries, not the 22 paired comparisons used for independent validation selection in §8. `A` denotes the full recorded grid and `B1` the same grid without its final point; both are analysis variants, not the two representation arms.

domain	metric family	condition–metric entries	V_λ median (A: full grid)	V_λ median (B1: final point removed)
NLP-AG	rank	16	0.0097	0.0074
NLP-AG	threshold	16	0.0325	0.0185
NLP-POS	threshold	24	0.2754	0.2312
VIS	threshold	24	0.2794	0.0973

entries. They are not the counts for the wider condition set in §6. The ratio between the archived cross-domain and music rank-based medians is 0.51 (0.0097 versus 0.0192); this arithmetic comparison does not establish a domain effect or a matched-protocol replication. The results document regularization dependence in the examined non-music settings without demonstrating broad generality.

The permutation-control summary records 40 comparisons (20 conditions × 2 metrics) across two domains with a usable criterion: 24 NLP-POS comparisons, of which all 24 pass, and 16 vision comparisons, of which 10 pass and 6 fail. The six failures are $deep_{\text{deep}}$ at `n = 2,000` and `n = 20,000` and $hand8_{\text{vs}}$_deep at `n = 20,000`, each failing under both metrics, and in all six the real effect lies below the permutation maximum; the criterion is therefore not usable in those large-`n` vision settings. In POS all comparisons pass, including the largest condition (`n = 204,192`,

tokens) at 3.9× its permutation maximum, so large n alone does not prevent detection. In POS, n counts tokens rather than tracks, so these sample sizes are not directly comparable.

Two boundaries of scope are stated explicitly. First, a separate archived NLP-AG batch of 16 comparisons records 2 passes and 14 failures and is not pooled with the 40, for two reasons. It uses a different probe specification, carrying a class-prior bias term that reproduces an archived $fit_{\text{ridge}} + predict$ convention and not the shared-coefficient solver used above, and compares layer and pretraining variants within AG News instead of the main batch's representation pairs. Its failures also come from the same mechanism as the vision ones, and are no different verdict: AG effects are the smallest in the archive (V_λ medians 0.0097 rank and 0.0325 threshold), and in 14 of 16 cases the real statistic sits at or below the permutation mean, so the criterion cannot separate them from the null. Pooling them would read as a 56-comparison tally with a 36% failure rate; they are reported apart because they are not one experiment. Second, no music comparison enters either paired batch. Three archived single-arm music checks stay outside those counts. A CLAP sweep over the dark twelve-label configuration permutes labels five times at each of nine coefficients: the permutation mean stays within 0.019 of chance at the eight coefficients $\alpha \leq 3 \times 10^5$ and falls to 0.462 at $\alpha = 10^6$, where the realised score is 0.798 — a failure of the centring criterion, not of separation: it exceeds the permutation maximum there by 0.28, and the sweep shares only $\alpha = 10^{-3}$ with the primary grid. Single-arm DEAM-valence checks fail the gain criterion for both, the permutation gain exceeding the realised gain (0.562 versus 0.206 for CLAP; 1.206 versus 0.456 for MERT-95M); single-arm Emotify GEMS-mean checks pass for both (realised gains 0.392 and 0.547 against permutation maxima 0.006 and 0.013). These are single-arm label-dependence checks; they supply no paired music comparison and are not pooled with the forty. The failures remain part of the evidence and prevent an unqualified claim that all controls pass.

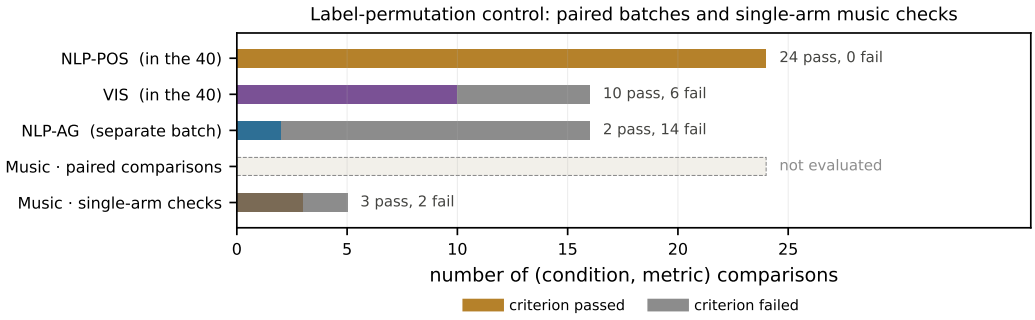

Figure 6: Domain coverage of the label-permutation control. The 40 evaluated comparisons are 24 NLP-POS and 16 vision entries. A separate archived NLP-AG batch of 16 comparisons, run on a different probe specification and the smallest archived effects, is not pooled with them. Paired music comparisons are absent; the single-arm music checks of Appendix A.2 cover MTG, DEAM and Emotify. Filled segments count comparisons meeting the criterion; grey segments count failures.

Four figures provide supporting diagnostics. Figure 7 displays df/d over the sampled range, from approximately 1.0 to 1.2×10^{-3} (about 800×), in the reported three-domain analysis. The smooth ridge diagnostic is not evidence that every empirical performance metric varies smoothly or that all possible regimes have been excluded. Figure 8 shows the MTG 56-label eighteen-point extended grid, with maxima at CLAP $\alpha = 3 \times 10^5$ and MERT-95M $\alpha = 10^5$ beyond the previous upper boundary $\alpha = 3000$; these are maxima on the extended tested grid, not proven continuous optima. Figure 9 relates reference-to-grid-maximum improvements to d/n. Figure 10 reports retrospective headroom relative to the reference score. Since headroom is defined using the observed maximum, it is not an independent predictor of recoverable performance or a causal explanation for a nonsignificant result.

A.3 Independent acceptable choices and common-set sizes

The shared intersection omits pairs in which the two arms use different acceptable coefficients, so we reanalyse the saved validation and refitted test curves for all 22 cross-domain comparisons,

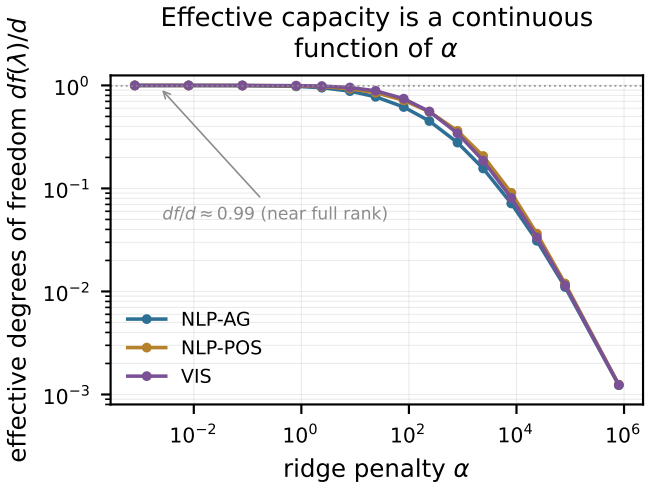

Figure 7: Effective-degrees-of-freedom diagnostic df/d over the sampled coefficient range in three domains, from approximately 1.0 to 1.2×10^{-3}. This characterizes ridge shrinkage under the stated coordinate convention.

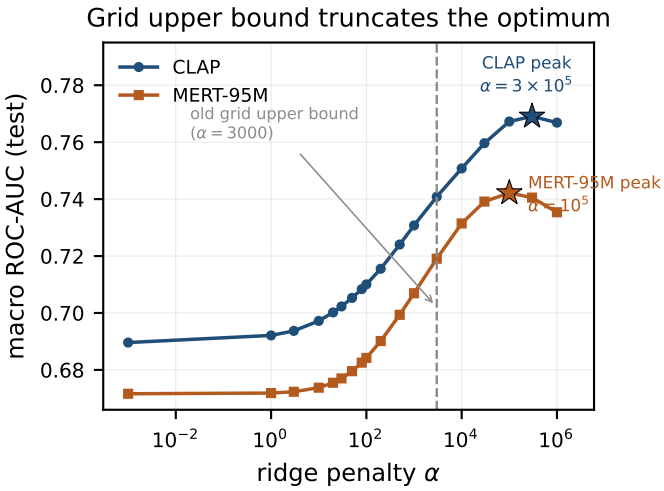

Figure 8: Extended ridge sweep with observed grid maxima at CLAP $\alpha = 3\times10^5$ and MERT-95M $\alpha = 10^5$, beyond the earlier grid boundary $\alpha = 3000$. These are sampled-grid maxima, not established continuous optima.

computing J^{π}_{ε} at the same three tolerances. This adds no model fitting: it is a descriptive summary of existing results, not a pre-registered analysis or an uncertainty interval. The accuracy-gap envelopes have median width 0.00815, 0.01153 and 0.03200 at the three tolerances, maximum width 0.08440, 0.08440 and 0.09330, and 2, 2 and 3 of 22 envelopes carry both signs.

Two AG News pretraining comparisons illustrate information absent from the common-set range. At $\varepsilon = 0.005$ each arm has three acceptable coefficients, but the shared intersection contains only one; the independently acceptable test-accuracy envelope is `[-0.00934, +0.01961]` for the 1,000-example pool and `[-0.01553, +0.02224]` for the 400-example pool. Each common-set range is exactly zero, yet each product set contains candidate pairs with positive and negative test-gap estimates. A singleton common intersection therefore does not certify a stable direction under independently acceptable choices. These are fixed-test point estimates that include choices other than the exact validation argmax: they establish no statistically supported ranking reversal,

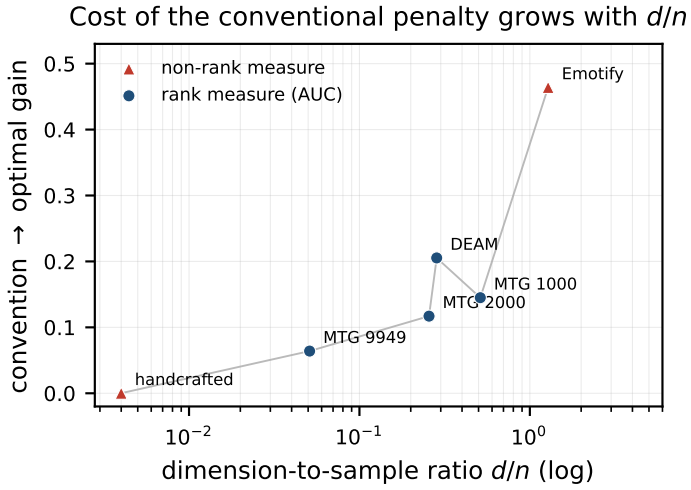

Figure 9: Reference-to-grid-maximum improvement versus d/n. Values retain their task and metric definitions; the relationship is descriptive.

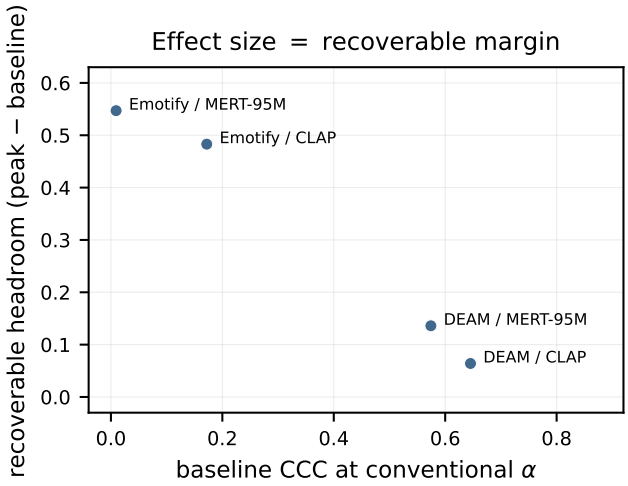

Figure 10: Retrospective headroom, defined as the observed grid maximum minus the reference score, versus reference performance. The panel title states this construction; it is not a prospective prediction or causal explanation.

and they do not authorize selecting the pair with the preferred test outcome. At $\varepsilon = 0.005$ the other twenty accuracy envelopes have a strictly consistent sign on their evaluated product sets, so the result identifies a limitation of common-set summaries rather than declaring every selected comparison unreliable.

B Metric Dependence of the Selected Penalty

Archived diagnostics referenced from §5. These readings are context for the primary interval rather than support for it; they use the same archived runs and add no fitting.

The coefficient that maximizes a score is not intrinsic to a representation; it moves with the metric and the protocol. CLAP's sampled peaks fall at $\alpha = 300$ for R^2 and 100 for the other three metrics, MERT-95M's at 1000, 300, 1000, 1000; between-backbone peak ratios run from 3 to 10 with metric

Table 10: Common-set sizes, with the actual collection and counting unit identified. The constructed batch contains twelve MTG entries and one DEAM entry. The native multi-metric music batch is not a set of NLP/vision conditions: its ten entries are two MTG AUC grid variants, two DEAM metrics, and six Emotify condition–metric entries. Repeated partitions, metrics, tolerances, and grids are not independent studies.

analysis collection	condition–metric entries	$\varepsilon = 0.005$ (empty / single / multi)	$\varepsilon = 0.01$ (empty / single / multi)	$\varepsilon = 0.02$ (empty / single / multi)
constructed-validation music batch (12 MTG + 1 DEAM)	13	1 / 6 / 6	0 / 1 / 12	0 / 1 / 12
matched native-validation MTG batch	12	5 / 1 / 6	1 / 4 / 7	0 / 2 / 10
native-validation / dataset-specific multi-metric music batch	10	4 / 6 / 0	2 / 8 / 0	1 / 7 / 2
repeated music folds and test subsets	16	2 / 14 / 0	2 / 14 / 0	1 / 11 / 4

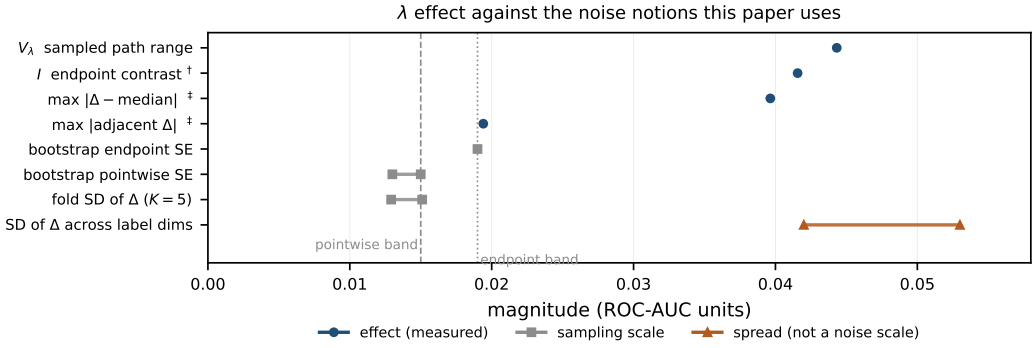

Figure 11: Primary-grid statistics and descriptive reference scales in ROC-AUC units. Circles show observed grid statistics, squares show reference scales, and the triangle shows across-label heterogeneity. Its vertical reference lines are not confidence-band boundaries or significance thresholds.

and tie-breaking, spread across specified choices and silent about sampling instability. These DEAM continuous-label peaks are not comparable with the $\alpha = 6.0 \times 10^4$ of Figure 1(a) or the $3{\times}10^5$ of Figure 8, which come from the MTG primary twelve-label ten-point grid and the MTG 56-label eighteen-point extended grid respectively.

The 56-label configuration's macro ROC-AUC row (Appendix A.1) comes from a separately defined S9 list that omits `ambiental` and is therefore no superset of the primary twelve, and a pre-specified full-label plan does not retroactively make the post-inspection contrast confirmatory.

The spectral ratios at the reference penalty are reported in §5 and indicate weak aggregate shrinkage in that diagnostic run, not a representation-invariant default. Because the third diagnostic equals twice the first minus the second, their agreement is not three independent pieces of evidence.

Single-arm ranges answer a different question from pairwise gap sensitivity. Reference-to-grid-maximum improvements span +0.00014 to +0.464 across six conditions on three datasets; the endpoints come from non-rank metrics and share no common scale with AUC changes. The archived published specification sits `0.079--0.097` ROC-AUC below its grid maximum. All of these compare against observed maxima retrospectively and promise nothing that validation selection would deliver.

The tested scaling hypotheses did not provide a consistent cross-condition description of the observed peak locations (Appendix F). We therefore use the spectral diagnostics and retrospective headroom as context, not as predictive or causal laws.

On the DEAM continuous-label conditions, with fold construction, data, probe, and candidate grid fixed, changing the reported metric gives the sampled peak locations in Table 11. The final column is the minimum ratio between a non-CCC peak coefficient and the CCC peak coefficient on the denser grid; it is not the minimum pairwise ratio among all four metrics.

Table 11: Metric-specific coarse-grid peak coefficients and denser-grid ratios. The four metric columns and $\log_{10}$ span describe the coarse grid; only the final column gives the smallest other-metric/CCC peak-coefficient ratio on the denser grid.

backbone	R^2	CCC	ROC-AUC	PR-AUC	$\log_{10}$ span	minimum other-metric/CCC ratio on denser grid
CLAP	300	100	100	100	0.477	2.37×
MERT-95M	1000	300	1000	1000	0.523	1.78×

With denser coefficient spacing (approximately 3.3× to 1.3×), the peak locations are `300 / 71.1 / 168.7 / 168.7` for CLAP and `948.7 / 400.1 / 1265.1 / 711.4` for MERT-95M, ordered as R^2 / CCC / ROC-AUC / PR-AUC. CCC has a lower peak coefficient than the other metrics on this denser grid, with minimum other-metric/CCC ratios of 2.37× and 1.78×. The maximum within-CLAP ratio increases from 3.0× to 4.22×. The metric dependence is therefore not confined to the coarser sampling used here, although neither continuous peak locations nor an unchanged total ordering of all metrics is established. Increasing the grouped-fold count from 5 to 10 leaves the reported coarse-grid maximizers unchanged; equality of grid maximizers does not imply equality of the underlying curves. Between-backbone peak ratios are 3.33× / 3.00× / 10.0× / 10.0× across the four metrics, with the two 10.0× values depending on tie-breaking. We report these as protocol-specific ratios, not as a universal ratio or a sampling-stability statistic.

The grouped-fold comparison also examined scores along the grid rather than only the maximizing coefficient. The archive records a weak-regularization exception for MERT-95M under R^2: the difference between the five- and ten-fold mean scores exceeds the reported cross-fold standard deviation there, even though the maximizing grid point is unchanged. This is a descriptive discrepancy, not an equivalence or difference test. It prevents an unchanged sampled maximizer from being presented as evidence that the entire score curve is unchanged.

Figure 12 displays the metric-specific coefficients. On the coarse grid, CLAP's CCC, ROC-AUC, and PR-AUC peaks coincide, whereas MERT-95M's CCC peak lies below its other three peaks. The separate loss-family check records single-arm ridge and logistic peak-to-trough changes of +0.0642 and +0.0858. At the reference penalty, the logistic score is 0.7134 and the ridge score is 0.7371. This supports single-arm sensitivity under both tested loss families, not a replicated paired-gap or ranking-reversal result.

C Paired Full-Grid vs Restricted Swings

Table 12 pairs matched MTG conditions across constructed- and native-validation rounds, with each condition's test set held fixed. Both the validation source and the associated fitting set change. Equal n_{test} values alone would not establish matched samples; the pairing follows the recorded condition identities. The entries describe the joint protocol comparison rather than an isolated effect of validation construction.

An entry of exactly 0.00000 denotes a singleton intersection, whose range is zero by construction rather than evidence of a flat test-gap path. A dash denotes an undefined restricted range when the common set is empty, not a zero range. The ten- and 21-point grids are non-nesting; their observed ranges do not establish numerical convergence.

D RAPTOR-Style Procedural Analysis

Datasets and representation pairs. The nine UD English-EWT conditions compare BERT-base-uncased layer 6 with layer 12, character TF-IDF/SVD-512, or character TF-IDF/SVD-8, each at pools of 445, 1,000 and 2,011 tokens. BERT features have 768 dimensions. The four AG News conditions compare English BERT layers 6 and 12, or `bert-base-uncased` and

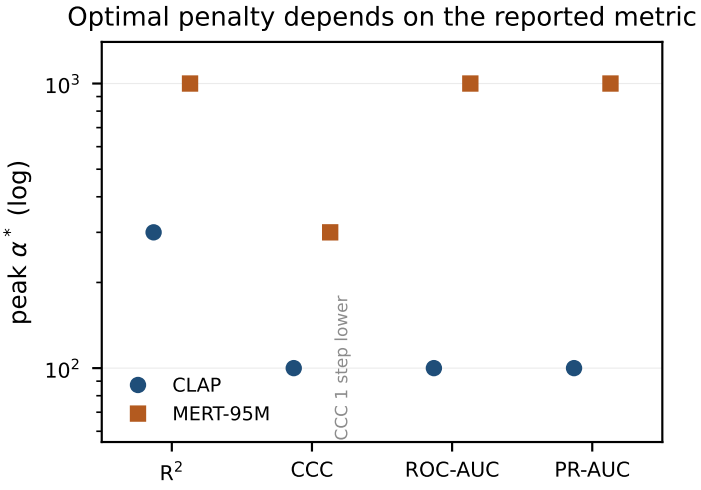

Figure 12: Selected coefficient by reported metric for both backbones. Coarse-grid ties and metric-specific differences are retained; the CCC coefficient is not strictly below every other coefficient in both coarse-grid series.

Table 12: Paired full-grid versus restricted swings. Each matched condition is paired across the constructed- and native-validation rounds with its test set held fixed, so both the validation source and the associated fitting set change. (c)/(n) mark the round; V_{full} is the swing over the full admissible set, V_{cmn} the swing over the common set at $\varepsilon = 0.01$, and "empty" records an empty restricted set at $\varepsilon = 0.005$.

condition	V_{full} (c)	V_{cmn} (c)	V_{full} (n)	V_{cmn} (n)	empty (c)	empty (n)
main, 10-point	0.04874	0.01639	0.04432	0.00000	non-empty	empty
main, 21-point	0.04825	0.01143	0.04490	0.00000	non-empty	empty
layer5 vs layer8	0.09028	0.00176	0.10376	0.00083	non-empty	non-empty
layer6 vs layer8	0.09456	0.00172	0.08773	0.00022	non-empty	non-empty
layer5 vs layer6	0.03017	0.00004	0.05898	0.00105	non-empty	non-empty
MERT-330M vs MuQ	0.04662	0.02004	0.04807	0.00518	non-empty	non-empty
window 10 s vs 30 s	0.04687	0.00142	0.14251	0.00153	non-empty	non-empty
MERT-95M (10 s) vs MuQ	0.04410	0.03209	0.04758	0.00000	non-empty	non-empty
CLAP vs MuQ	0.01285	0.00253	0.01434	0.00193	non-empty	non-empty
handcrafted vs MERT-95M	0.07440	0.04367	0.06836	0.02113	empty	empty
noisy10 CLAP vs MERT-95M	0.04795	0.01578	0.04602	0.00000	non-empty	empty
noisy25 CLAP vs MERT-95M	0.04889	0.01462	0.04452	—	non-empty	empty

`bert-base-multilingual-uncased` at layer 12, each with 400 and 1,000 documents. The nine CIFAR-10 conditions compare ResNet-18 (`resnet18.a1_in1k`, 512 dimensions) with ViT-Tiny (`vit_tiny_patch16_224.augreg_in21k`, 192), or pixel PCA-8/PCA-512 with ResNet-18, each with 400, 1,000 and 2,000 images. Recorded test sizes are respectively 25,093 tokens, 7,600 documents and 10,000 images.

Pool sizes precede the internal split (requested validation fraction 0.2; seed 20260914); every condition uses 15 ordered solver coefficients. Realized validation sets hold 79–400 examples: one accuracy step spans 0.0025–0.0127 and the validation standard error 0.012–0.056, so $\varepsilon = 0.005$–0.02 spans at most 1.7 such errors. The POS loader filters sentences exceeding 128 BERT tokens and fits word-boundary character 2–4-gram TF-IDF followed by truncated SVD. Image PCA uses flattened pixels scaled to $[0, 1]$. The loader constructs these baselines before selecting the small probe pools. The POS task inventory contains 17 tags; at 445 tokens the stored arm-level output

count is 16, following $C = 1 + \max(y_{\text{pool}})$ in the fitting code. This sample-dependent output count is distinct from the task inventory.

The cross-domain analysis adopts a RAPTOR-style validation-selection and refitting procedure (Gao et al., 2026), using our own stored candidate grids and ridge implementation. It is not a reproduction of the full RAPTOR pipeline, application, grid, or parameterization. Validation accuracy selects the coefficient; all candidates are refitted on training plus validation with the same numerical coefficients, and the test oracle is evaluated separately for each reported metric. The single-arm summaries in Table 13 use the same 22 conditions as §8.

Table 13: Single-arm summaries under the RAPTOR-style procedure. Values cover the same 22 conditions and 44 arm-level fits as §8. Shortfalls are taken against each arm's observed grid maximum and involve no resampling; the coefficient-ratio and direction-stability rows are descriptive summaries of the stored selected coefficients.

quantity	value
conditions	22
arms	44
single-arm shortfall, median	0.0036
single-arm shortfall, maximum	0.0457
share with shortfall ≤ 0.01	73%
share with shortfall = 0	30%
span of the stored selected solver coefficients	approximately 7 orders of magnitude
within-condition two-arm ratio of selected nominal coefficients	median 10×, max $10^6\times$
reported probe-vector direction-stability statistic (penalty reselected)	median 0.848, range [0.664, 0.991]

The median single-arm accuracy shortfall is 0.0036 and its observed maximum is 0.0457; 32 of 44 arms have shortfall at most 0.01 and 13 have zero shortfall. Selected indices occupy seven grid positions. The coefficient span and within-pair ratios describe stored solver values, not within-arm sampling instability; the explicit grids and selected indices are stored in the per-condition artifacts.

The auxiliary direction statistic averages absolute pairwise cosines of normalized probe vectors over `K = 20` perturbation-and-reselection repetitions, with 20% removal. In the recorded per-class convention, vectors are normalized within class before averaging: $Robust = [2/(K(K-1))] \sum_{r<s} \text{mean}_c\, |\langle v_{r,c}, v_{s,c} \rangle|$. It concerns fitted probe direction, not the sign of the A–B performance gap. The supplied scalar values are retained as an auxiliary diagnostic; their precise handling of any zero-norm class vectors remains part of the archived implementation, not a ranking-stability guarantee.

For score reconstruction, selected indices and score keys are matched to each arm's explicit stored solver grid. All 22 pairs have matching within-pair grids and split fingerprints, and recomputation reproduces the reported selected scores, metric-specific maxima, shortfalls and paired offsets.

The grid uses the planned training count $n_{\text{grid}} = \text{round}((1-v)n_{\text{pool}})$ and $\alpha_k = (n_{\text{grid}}/2)\lambda_{\text{aux},k}$; class-wise rounding can change the realized fitting count. The numerical solver coefficients remain fixed on refitting. The supplementary `metadata/parameter_grid_semantics.md` records planned and realized counts, legacy field meanings and small cross-condition grid differences. Figure 13 plots selected coefficients against empirical single-arm shortfalls.

E VERIFICATION

The implementation checks are listed in §4. The reordering invariance is distinct from the label-permutation control. An independently implemented reference calculation comparing the closed-form solution with $sklearn.linear_{\text{model}}.Ridge$ agrees to 3.56×10^{-16} relative in predictions and 2.04×10^{-16} in the objective, against predeclared tolerances of 10^{-8} and 10^{-10}. The archived analysis contains 178 passing assertions, including per-condition recomputation from the recorded paths. A deliberately corrupted replicate causes 36 assertions to fail, demonstrating sensitivity to that corruption. Replay reproduces the archived scores bitwise in the recorded environment. These

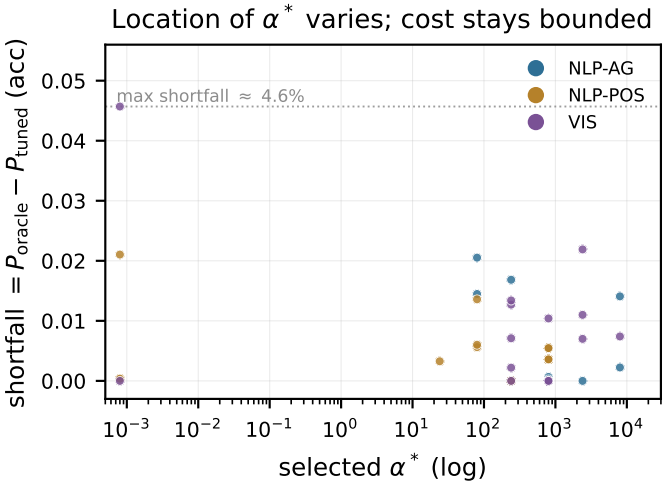

Figure 13: Selected coefficient versus empirical accuracy shortfall across 44 arms. The non-negative ordinate is $P_{\text{oracle}} - P_{\text{tuned}}$. The observed maximum, 0.0457 (4.57 percentage points), is a finite-sample maximum, not a population bound; "bounded" in the panel title has no such inferential meaning.

checks support the numerical implementation and traceability of the tested calculations; they do not establish the correctness of every scientific interpretation or independent external reproduction.

F Withdrawn Claims and Scope

Two auxiliary analyses delimit the scope. Under a pre-specified auxiliary hypothesis and decision criteria, prediction agreement for a fixed representation was expected to decrease under sample scarcity. A bootstrap-CI analysis and a controlled mutually exclusive-subset design gave inconsistent support, so the hypothesis was withdrawn under those criteria. An earlier cross-backbone agreement result was also withdrawn after reimplementation: two separate implementations agree with each other ($\rho_{\text{same}} \approx 0.93 > \rho_{\text{cross}} \approx 0.52$) and disagree with the original run. These records document excluded interpretations; neither is used as positive evidence for the paper's main claims.

Several proposed scaling relationships also failed to provide a consistent description across the tested conditions. Table 14 records the observations that led us not to adopt them as general characterizations. The observations are not universal impossibility results for other data or parameterizations, and the regime-related row concerns the tested grid-boundary interpretation.

Table 14: Characterizations considered and not adopted, with the observation that excluded each. These rows record interpretations the archive does not support. They are not universal impossibility results for other data or parameterizations, and none is used as positive evidence for the paper's claims.

characterization not adopted	observation motivating the limitation
$\lambda* \propto n$	the same $\lambda*$ observed across datasets with different $\mathtt{n}$
$\lambda* \propto d$	monotone but not linear
$\lambda* = c \cdot \lambda_{1/2}$	49× range within MTG
$\kappa \approx 0.3$ universal	contradicted on Emotify
$\mathtt{d/n}$ or task type sets a discrete regime	8/8 interior solutions after grid extension

These observations limit the explanations advanced here. They do not establish an alternative universal scaling law, and agreement or disagreement on a finite candidate grid is not a proof about continuous optima.

G Implementation Details

The reported environment is PyTorch 2.6.0+cu124, scikit-learn 1.9.0, and librosa 1.0.0. These are the recorded version declarations, not a reconstructed environment in this revision. The principal probe is centered, column-standardized ridge with an unpenalized intercept. The primary ten-point solver grid is `logspace(-3, 7, 10)`; the separate 21-point grid is non-nesting. The primary path uses 327 test-artist clusters and 400 paired resamples; the alternative Figure 5 configuration uses 10,000 resamples. S-42c uses 500 **track-level** resamples for each of its two separate perturbation designs. Primary encoders remain frozen, with the recorded 10 s windows and 5 s hop.

The cross-domain procedure selects by validation accuracy and retains the selected numerical solver coefficient for training-plus-validation refitting. Every metric-specific test oracle is computed over the same refitted candidate family. The grid-construction count is $int(round((1 - val_{\text{frac}}) \cdot n_{\text{pool}}))$, not necessarily the realized training count after stratification. Exact stored solver grids and selected indices define the reported results; the legacy $alpha_{\text{conversion_n}}$ field records actual training rows and is not used to rebuild the grid (Appendix D). In S-42b, the recorded training cap is distinguished from the realized fitting count. Auxiliary pre-run plans apply only to their named analyses and do not retroactively make the exploratory main contrast or the new saved-score envelopes confirmatory.

H Archived Normalized-Complexity Alignment

An existing auxiliary analysis varies a common target $\rho = df(\alpha)/d$, solving for each arm's own coefficient. The declared target list is {`0.99, 0.95, 0.90, 0.80, 0.60, 0.40, 0.20, 0.05`}; only targets reachable by both arms in the stated condition are retained. The archived results contain 32 condition–sample-size entries and 217 retained target rows. The df_{A} and df_{B} fields in this particular artifact store **normalized** `df/d`, not absolute degrees of freedom. Table 15 illustrates one vision condition, $deep_{\text{deep}}$, with 400 fitting examples and feature dimensions 512 and 192.

Table 15: One vision condition under a common normalized-complexity target. $\rho = df(\alpha)/d$ is the common target both arms solve for; the columns give each arm's own solver coefficient and the resulting test-accuracy gap `A − B`. Only targets reachable by both arms in the stated condition are retained.

common target ρ	solver α_{A}	solver α_{B}	test-accuracy gap A − B
0.60	24.08974	74.17812	+0.04580
0.40	135.59679	235.85139	+0.10730
0.20	693.96306	908.30293	+0.14540
0.05	5658.49969	6213.48245	+0.20060

All four retained target points have positive gaps. At the separate common-coefficient anchors in the same archived condition, the gaps are `−0.00610` at $\alpha = 1$ and `−0.02910` at $\alpha = 0.01$. Those anchors do not match `df/d` between arms. The comparison protocol and sampled region have both changed, so this is not evidence that matching complexity causally removes all reversals. It demonstrates why a nominal-coefficient path, a normalized-complexity path, and their respective attainable target ranges must be stated separately. A subsequent common-target restriction across sample sizes does not make the underlying samples or tasks independent.

The archived runs contain no matched-complexity analysis for the primary MTG configuration, so one was recomputed from the stored per-arm test curves under the recorded definition: a common target $\rho = df/d$, each arm solved for its own coefficient. Table 16 lists the resulting nine-rung ladder, extending the archived target list by one rung at $\rho = 0.001$. Over the shared coverage window (α_{A} from 4.3×10^{-3} to 8.1×10^{4}) the matched path gives an endpoint contrast of $+0.00864$ against the nominal path's $+0.00793$ over the same window; at $\rho = 0.001$ ($\alpha_{\text{A}} = 9.72 \times 10^{6}$, $\alpha_{\text{B}} = 9.64 \times 10^{6}$, both arms at $df/d \approx 9.7 \times 10^{-4}$) the matched contrast is $+0.04171$ against the nominal $+0.04156$. Once coverage is aligned the two conventions agree, and the matched path's full range, 0.05561, exceeds the nominal path's 0.04432. The gap stays positive at every rung; the sign reversal in the vision condition of Table 15 does not occur here. Across the 22 archived cross-domain comparisons the matched path leaves four comparisons with sign-changing gaps.

Table 16: Primary MTG configuration under a common normalized-complexity target, recomputed from the stored per-arm test curves. $\rho = df/d$ is the common target both arms solve for; the last column reads the shared-coefficient path at α_{A} by log-linear interpolation. The final rung extends the archived target list.

common target ρ	solver α_{A}	solver α_{B}	matched gap	nominal Δ at α_{A}
0.99	0.0043008	2.3406	+0.032978	+0.033542
0.95	0.28854	14.931	+0.031468	+0.034684
0.90	1.7944	35.301	+0.029427	+0.035817
0.80	11.881	96.679	+0.026230	+0.037222
0.60	101.09	385.74	+0.021347	+0.034984
0.40	547.51	1416.8	+0.019077	+0.031871
0.20	4180.5	7588.1	+0.025334	+0.030525
0.05	80580.6	90049.8	+0.041621	+0.041467
0.001	9.7227×10^6	9.6443×10^6	+0.074684	+0.074599

I Primary MTG Pointwise Summaries

The primary pointwise table (Table 1) sits in §5 as the tabular form of Figure 1; its values come from mtg_{main}_NATIVEval_10pt in the native-validation archive and are not read from the figure. The twelve labels are action, adventure, advertising, ambiental, background, ballad, calm, children, christmas, commercial, cool, and corporate. Fitting uses 9,949 training tracks; validation has 3,802 tracks; evaluation has 4,231 tracks in 327 recorded artist clusters. Intervals are nominal pointwise percentile intervals from 400 paired test-cluster resamples.

The endpoint contrast and range are calculated from these gap estimates; its endpoint interval is computed from paired endpoint differences within resamples, not by subtracting the endpoints of the two pointwise intervals.

J Setup, Protocol, and Metric Details

This appendix holds the protocol and definitional detail that the main text states only in condensed form. It carries no result that is not already reported in the main text or in the appendices above.

Spectral summaries. To state the three archived spectral diagnostics without a matrix-dimension ambiguity, write $K_\alpha = (X_{\mathrm{c}}^T X_{\mathrm{c}} + \alpha I_{\mathrm{d}})^{-1} X_{\mathrm{c}}^T X_{\mathrm{c}}$ and $h_{\mathrm{k}} = \sigma_{\mathrm{k}}^2/(\sigma_{\mathrm{k}}^2 + \alpha)$. The reported ratios are $\mathrm{tr}(K_\alpha)/d = \sum_k h_{\mathrm{k}}/d$, $\mathrm{tr}(K_\alpha^2)/d = \sum_k h_{\mathrm{k}}^2/d$, and $[d - \mathrm{tr}((I_{\mathrm{d}} - K_\alpha)^2)]/d = \sum_k (2h_{\mathrm{k}} - h_{\mathrm{k}}^2)/d$. Only the first is the usual fitted-value effective-degrees-of-freedom trace; the others are related shrinkage summaries, not independent estimates of a representation-invariant quantity. Matching `df/d` is distinct from matching absolute `df` or rank-normalized `df`. When λ denotes the solver coefficient α, dividing the normal equations by a declared sample count gives α/n; this does not remove spectral or preprocessing differences.

Selection surface and shortfall summaries. The corresponding comparison surface is $G^\pi(\lambda_{\mathrm{A}}, \lambda_{\mathrm{B}}) = T_{\mathrm{A}}^\pi(\lambda_{\mathrm{A}}) - T_{\mathrm{B}}^\pi(\lambda_{\mathrm{B}})$; where fitting policies agree, the shared path is its diagonal $G^\pi(\lambda, \lambda)$, while independent selection lands off that diagonal. Because the shared sweep uses training-only fits and the selection analysis refits, the two also differ in fitting policy, so neither result is a compression of one unchanged path. Shortfalls are summarized by their fraction at zero, medians, maxima, and empirical upper quantiles as each analysis requires. The oracle reads the test set retrospectively and is not an admissible selection rule, so its shortfalls bound no future selection error, and optimism in the separate empirical maxima makes the paired offset neither conservative nor anti-conservative. An observed maximum, a resampling quantile, and a population bound stay distinct throughout.

Metric identities. Results retain their metric and condition identities. ROC-AUC and PR-AUC summarize ranking, accuracy and macro-F1 use classification decisions, and R^2 and CCC assess different properties of continuous predictions; metric-family summaries therefore pool quantities that share no common scale. A sign change in the sampled point estimates, an endpoint contrast, and a supported ranking reversal are three different claims, and the paper makes only the first two.

Normal equations and loss scaling. The normal equations correspond to minimizing $\|X_{\mathrm{c}}W - Y_{\mathrm{c}}\|_{\mathrm{F}}^2 + \alpha\|W\|_{\mathrm{F}}^2$; this specifies the solver-coordinate convention and assumes nothing about how other implementations name their regularization argument. Averaging a loss over samples or outputs changes the conversion to α unless the penalty is scaled consistently. The closed-form solution removes stochastic optimization as a source of within-sweep differences, not numerical or preprocessing dependence.

Interval construction for the alternative configuration. Its displayed simultaneous band is the nominal Bonferroni percentile band using tail probability `0.05/(2L)` per side for `L = 10` points. An additional archived max-type band uses $c_{0.95} = \mathrm{quantile}_{0.95}[\max_j |\Delta_{b,j} - \hat{\Delta}_j|/SE_j]$ and endpoints $\hat{\Delta}_{\mathrm{j}} \pm c_{0.95}SE_{\mathrm{j}}$ (Hothorn et al., 2008). Both bands belong to the alternative configuration; neither is substituted for the primary configuration's uncertainty.

Resampling scope. The S-42c validation and test bootstraps resample individual tracks rather than artist clusters (§8); pooling them with the cluster-bootstrap path analysis would overstate what they capture about artist-level dependence. Overlapping folds also remain dependent (Bengio and Grandvalet, 2004), and small evaluation sets limit precision (Varoquaux, 2018). None of these procedures is a DeLong test. CCC denotes the concordance correlation coefficient (Lin, 1989).

Permutation-control bookkeeping. The control addresses label dependence, not invariance to information-preserving reparameterizations of the features; it is therefore different from a feature-duplication control, and neither role is inferred from the other. Degeneracy checks identified exploratory reversals that coincided with collapsed predictions, and those cases were withdrawn from substantive reversal claims. Full-grid summaries are distinguished from summaries with the last point removed. Small prediction variance alone does not show that all scores are tied or that rank information has disappeared, and points outside a validation-near-optimal set are not automatically degenerate. Appendix E describes implementation and provenance checks. The main interpretation remains conditional on the stated nominal coordinate and requires no unreported coordinate-invariance result.

Per-label AUC aggregation. Undefined per-label AUC values are omitted by the recorded finite-value macro aggregation, so the calculation does not force every resample to retain the same number of evaluable labels; per-resample valid-label counts are not available in the supplied summary artifacts.

Remaining per-condition settings. The MTG selection experiment fixes CLAP and 20 output labels, uses an 11-point grid and eight grouped partitions, and requests training-sample caps of 9,949 or 2,000. Actual fitting counts are the smaller of the cap and the available training pool, not guaranteed to equal the cap in every partition, and the nominal 60/20/20 allocation concerns groups rather than fixed numbers of tracks. S-42c uses the official split and separate track-level test and validation bootstraps with `B = 500` each.

MTG partition and resampling detail (§8). The analyses below are supporting single-arm measurements referenced from §8; they use the same saved artifacts and add no new fitting.

A further MTG analysis fixes CLAP, 20 output labels, and eleven solver coefficients {`1, 10, 30, 100, 300, 1000, 3000,` 10^4, 3×10^4, 10^5, 3×10^5}, drawing eight 60/20/20 grouped partitions from a pool recorded as 1,508 artist groups and sampling training rows without replacement up to the requested cap. The stored `n` labels are those caps; the summaries preserve neither every partition's realized track count nor its group count. Each shortfall is the maximum candidate test AUC minus the test AUC at the validation-AUC-selected coefficient, so these repetitions say nothing about variability of a two-representation comparison. Table 17 reports them.

On the official-split CLAP configuration, S-42c perturbs evaluation and selection independently. Both designs use `B = 500`, seed 424242, the same eleven coefficients, 3,802 validation tracks, and 4,231 test tracks. **T-bootstrap** resamples individual test rows, holds the coefficient chosen on the complete validation set, and recomputes the empirical test oracle in each replicate; **V-bootstrap** resamples individual validation rows, reselects the coefficient, and evaluates against the fixed complete-test oracle. Both are track-level rather than artist-cluster bootstraps, with models held fixed while the scores are recomputed. Table 18 reports the two designs.

Table 17: MTG single-arm shortfall across eight partitions. Shortfalls are taken relative to the observed grid maximum, so they are not population bounds.

requested training cap	partitions	zero observed shortfall	median observed shortfall	maximum observed shortfall	selected coefficients (counts)
9,949	8	6 / 8	0.000000	0.001585	3×10^4 (1), 10^5 (6), 3×10^5 (1)
2,000	8	4 / 8	0.000128	0.004555	3×10^4 (6), 10^5 (2)

Table 18: Two resampling designs for the CLAP shortfall. The designs resample different units and are not interchangeable.

resampling design	empirical zero-shortfall fraction	median shortfall	empirical 95th percentile	observed maximum
test rows; validation selection fixed	7 / 500	0.001415	0.002321	0.003042
validation rows; test scores fixed	114 / 500	0.001410	0.001410	0.001410

K Null Calibration for the Primary Path

This appendix reports the primary configuration's range under grid trimming and under a constant-gap null. Everything here is computed from the same 10,000-replicate artist-cluster bootstrap as the endpoint interval of §5 (327 clusters, seed 20260926, predictions held fixed); no new model is fitted, and the `dark` configuration is not used.

Grid trimming. V_λ ranges over a declared ten-point grid and therefore inherits its extent. Removing the final coefficient gives $V_\lambda^{(\mathrm{B1})} = 0.02818$; removing the final two gives $V_\lambda^{(\mathrm{B2})} = 0.00877$. The final two coefficients account for 0.03555 of the full-grid 0.04432 (80%) and the final one for 0.01614 (36%). The arms remain far from degenerate at the largest coefficient: their absolute test ROC-AUC there is 0.795 and 0.720, the former exceeding the latter's grid-wide maximum by 0.029. Table 19 collects the three conventions.

Table 19: Grid-trim sensitivity of the primary sampled range. A: all ten coefficients; B1: final coefficient removed; B2: final two removed. The share column attributes the full-grid range to the removed tail.

convention	V_λ	ratio to A	tail attribution
A (ten coefficients)	0.04432	1.00	—
B1 (nine coefficients)	0.02818	A/B1 = 1.57	leave-one-out 36%
B2 (eight coefficients)	0.00877	A/B2 = 5.05	leave-two-out 80%

Constant-gap null. A range is an order statistic; comparing it with a pointwise standard error understates its null variability. Under the null hypothesis of a coefficient-independent gap we centre each bootstrap draw per coefficient, $\Delta_b(\alpha) - \hat{\Delta}(\alpha)$, and record the range of the centred path in each replicate. Table 20 reports observed ranges against this null. The full-grid range sits at the 94.7th null percentile ($p = 0.053$); the three coefficients from 4.6×10^3 to 7.7×10^5 ($p = 0.012$) and the final two ($p = 0.005$) lie clearly in the tail, and the plateau segment $\alpha \le 4.6 \times 10^3$ does not ($p = 0.81$). Independence would put the ten-point null mean at $d_2(10) \times \mathrm{SE} \approx 3.08 \times 0.0137 = 0.042$; the measured null median is 0.0218, an effective d_2 of 1.59, because the ten gap estimates are strongly dependent (Figure 2b).

Bootstrap dependence. The ten gap estimates form two blocks: within $\alpha \le 27.8$ the median pairwise correlation is 0.991, within $\alpha \ge 359$ it is 0.717, and across the two blocks 0.192, with $r(\Delta(10^{-3}), \Delta(10^7)) = 0.093$ and $r(\Delta(7.7 \times 10^5), \Delta(10^7)) = 0.873$. The small eight-coefficient range and the near-zero endpoint correlation are consistent with one description: correlation decays

Table 20: Observed primary ranges against the constant-gap null (10,000 centred bootstrap paths). p is one-sided: the null probability of a range at least as large as observed. Only the post-peak rise from 6.0×10^4 to 10^7 is prespecified; the other subsets were chosen while inspecting the grid, so their values are exploratory.

subset	k	observed range	null median (95th pct.)	p
all ten coefficients	10	0.04432	0.02184 (0.04466)	0.0528
$\alpha \leq 5.99 \times 10^4$ (to the sampled peak)	8	0.00877	0.01652 (0.04115)	0.8127
final three (the post-peak rise)	3	0.03555	0.00883 (0.02487)	0.0028
4.6×10^3–7.7×10^5 (three points)	3	0.02818	0.00811 (0.02185)	0.0120
final two coefficients	2	0.01614	0.00423 (0.01094)	0.0049
plateau $\alpha \leq 4.6 \times 10^3$	7	0.00741	0.01520 (0.04045)	0.8065

with α separation, and a block-structured account fits the matrix where an exchangeable shared-plus-independent model does not.

Simultaneous bands. At $B = 10{,}000$ the ten-point Bonferroni band (tail probability $0.05/(2 \times 10)$ per side) and the max-type band ($c_{0.95} = 2.524$) both contain a constant gap curve: the largest lower bound minus the smallest upper bound is negative for each, as in the archived `dark` configuration. Raising the resample count does not change this. The band criterion is pointwise and therefore conservative for a range statistic: admitting a constant curve is not in conflict with the null percentile of the two-point range reported above. The endpoint interval at $B = 10{,}000$ is $[+0.00765, +0.07980]$; the bootstrap reflects evaluation-sample uncertainty conditional on the fitted probes and the split (§4), so the three intervals below are not independent random sources: $[+0.00499, +0.07935]$ (400 resamples, seed 20260926), $[+0.01257, +0.08194]$ (400 resamples, seed 20260918, archived validation-alignment run), and $[+0.00765, +0.07980]$ (10,000 resamples, seed 20260926). Of the ten adjacent-step contrasts I_{ij}, three exclude zero at $B = 10{,}000$: the final two steps and the endpoint pair, the same region the null analysis identifies.