THE LOW-RANK BRAIN: WHY FMRI ENCODING-MODEL PREDICTIONS CAN MISLEAD DECODING AND LOCALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Predictions from fMRI encoding models are increasingly analyzed as if they were measured brain activity: decoded, correlated, and compared across cortical regions. But a predicted response is a function of stimulus features, and its many vertices need not carry independent or regionally specific information. We examine this with two audio-to-fMRI encoders applied to 8,625 fifteen-second solo-piano clips spanning 12 composers and three stylistic eras. The predictions of the first, TRIBE v2, span 20,484 cortical vertices, yet 6, 9, and 19 standardized principal components explain 90%, 95%, and 99% of their variance. A linear ridge encoder, fit independently to one subject's natural-speech fMRI and applied out of domain, is lower-rank still (3, 6, and 16 components), suggesting that low rank is not specific to one encoder, although this exploratory comparison varies several factors at once. Era is linearly decodable from TRIBE's predictions and transfers to held-out composers under leave-one-composer-out (LOCO) cross-validation (balanced accuracy 0.469, chance 0.333; composer-permutation $p = 0.004$), whereas the ridge predictions do not decode it (0.280). Yet when era is decoded within each of 22 cortical divisions of the HCP-MMP1.0 atlas, no division beats a size-matched random-vertex null (max-statistic family-wise $p = 0.465$): the resulting map looks spatially organized, but no region carries more era information than an arbitrary vertex set of equal size. Encoding-model predictions can thus support decoding without supporting region-level localization. We recommend reporting the effective rank of predicted responses and testing region-level effects against size-matched random-vertex nulls before drawing spatial conclusions. Our claims concern downstream analyses of model predictions, not the validity of encoding models or of measured fMRI.

1 INTRODUCTION

Encoding models that predict brain responses from stimulus features are now central to computational neuroscience (Naselaris et al., 2011). A model is fit to map stimulus features onto measured fMRI, and its quality is judged by prediction accuracy on held-out responses. As these models improve, their predictions have acquired a second use: a trained encoder can emit a cortical response for stimuli that were never scanned, making predicted responses attractive for decoding, correlation analyses, and regional localization.

This second use raises a basic question about what can be inferred from predicted responses. A predicted response has one value per cortical vertex and inherits the anatomical labeling of the atlas used to train the encoder, so downstream analyses produce maps that are formally identical to those computed from measured fMRI. That correspondence does not imply that the predictions contain independent or regionally specific information. The predictions may instead occupy a small effective subspace. When many vertices are generated from a low-dimensional representation, different regions become different readouts of the same latent variables, and apparent regional differences can reflect sampling and representation structure rather than regional specialization.

We make this concrete with music. Using TRIBE v2 (d'Ascoli et al., 2026) (hereafter TRIBE), an audio-to-fMRI encoder, we predict cortical responses to 8,625 fifteen-second solo-piano clips

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.

spanning twelve composers and three stylistic eras, and ask of the predictions: can musical era be decoded from them, and if so, where? Era classification is a test case, not an end in itself. The task offers two natural controls: the audio features from which the predictions are generated, and a composer-level generalization test that distinguishes learning stylistic era from memorizing composer identity.

Three results organize the paper. First, the predicted responses are low-rank relative to their dimension: 6, 9, and 19 standardized components capture 90%, 95%, and 99% of the variance of a 20,484-vertex space. A second, independently fit ridge encoder, trained on an unrelated natural-speech fMRI dataset from a different subject, is lower-rank still. The comparison varies several factors at once and is exploratory, but it suggests that low rank is not a TRIBE-specific artifact. Second, era is decodable from TRIBE's predictions, yet no cortical parcel exceeds a size-matched random-vertex null after accounting for the 22 tested regions, so a visually organized cortical map does not establish localization. Third, era transfer to held-out composers exceeds a composer-permutation null for the predicted responses but not for the input audio representation; the difference in mean accuracy is not significant and changes sign across retained dimensionality, and we treat it as a non-result.

None of this bears on the validity of fMRI encoding models, which are judged on held-out predictive accuracy. Our target is a downstream practice: running decoding and localization analyses on encoding-model predictions as though those predictions shared the dimensionality and spatial structure of measured responses. We recommend two checks before such analyses: report the effective rank of the predicted responses, and test regional effects against a size-matched random-vertex null. Together they ask whether an apparent spatial pattern exceeds what the representation itself would produce from any vertex set of the same size.

2 Related work

Encoding models that predict fMRI from stimulus features are established in vision, language, and audition and are typically validated by prediction accuracy on held-out measured responses (Naselaris et al., 2011). TRIBE (d'Ascoli et al., 2026) follows this paradigm and was developed for the Algonauts 2025 challenge (Gifford et al., 2025). Its predictions can be decoded, correlated across conditions, and compared across cortical regions (d'Ascoli et al., 2026). We take the encoder as given and ask whether its predictions retain enough effective dimensionality for regional comparisons to support localization.

Effective dimensionality, measured by variance-threshold rank or the participation ratio, is a standard descriptor of neural population activity and of artificial-network features (Gao et al., 2017; Stringer et al., 2019). Measured cortical responses are themselves structured: in human auditory cortex, six components account for more than 80% of the noise-corrected voxel response variance to 165 natural sounds (Norman-Haignere et al., 2015). Low dimensionality is therefore not in itself a defect. We tie the descriptor to localization validity: when a representation's effective rank is small relative to its vertex count, every region is a different linear readout of the same few latent dimensions, which limits what a comparison across regions can establish.

Claims about the spatial structure of brain maps are commonly tested against nulls that preserve spatial autocorrelation, such as spherical spin permutations (Alexander-Bloch et al., 2018) and variogram-matched surrogate maps (Burt et al., 2020), compared by Markello & Mišić (2021) and implemented in neuromaps (Markello et al., 2022). These nulls ask whether two maps correspond beyond what their smoothness alone would produce. The size-matched random-vertex null used here asks a different question, namely whether a region carries more decodable information than an arbitrary set of vertices of equal size; the two are complementary (§5.3).

Era, style, and composer classification are long-standing tasks in music information retrieval, usually approached with acoustic descriptors such as timbral summaries, chroma, and MFCCs (Tzanetakis & Cook, 2002). We use era decoding only as a testbed and benchmark the predicted responses against such acoustic features; the contribution is methodological.

3 Methods

Appendix A gives implementation details omitted here, and Figure 3 the full pipeline.

3.1 Scope and object of analysis

Our object of study is the *output* of fMRI encoding models, that is, the cortical responses they predict from audio, and not measured fMRI. Every quantity we report for the piano stimuli is a deterministic function of the input audio; no scanner data were collected, no human participants are involved, and no institutional review was required. We restrict our claims to what those predictions do and do not support under downstream decoding and localization.

3.2 Encoding model

TRIBE (d'Ascoli et al., 2026) is a multimodal transformer fMRI encoding model that maps video, audio, and text onto predicted cortical fMRI on the fsaverage5 surface (20,484 vertices, $2 \times 10{,}242$). It was trained on the Algonauts 2025 (Gifford et al., 2025) naturalistic-movie data and is run at inference with `average_subjects=True`, so predictions correspond to a single averaged predicted subject. The audio pathway wraps a Wav2Vec2-BERT front end (Barrault et al., 2023). We drive only the audio pathway (§3.3).

3.3 Audio-only inference

TRIBE's convenience entry point runs a WhisperX (Bain et al., 2023) large-v3 speech-recognition stage. On instrumental piano, that stage hallucinates lyrics, and the spurious word events feed the model's text branch and contaminate its predictions. We therefore call the module-level `get_audio_and_text_events(df, audio_only=True)` directly, which skips speech recognition entirely. With no word events present, the loader drops the text and video extractors and `FmriEncoder.aggregate_features` substitutes zero tensors for the absent modalities, which is the model's own trained modality-dropout behavior; an assertion that no word events are present guarded every batch. We flag this because the default path fails on instrumental audio and, to our knowledge, would contaminate any naïve application of this model to music.

3.4 Stimuli

Stimuli are solo-piano performance recordings from MAESTRO (Hawthorne et al., 2019), cut into non-overlapping 15-second windows. The final set is 8,625 clips spanning 12 composers and 3 stylistic eras: Baroque (2 composers: Bach, Scarlatti; 1,389 clips), Classical (3: Beethoven, Haydn, Mozart; 1,791), and Romantic (7: Chopin, Schubert, Liszt, Rachmaninoff, Schumann, Scriabin, Brahms; 5,445). Debussy was removed before analysis as the sole member of an Impressionist/Modern class: with one composer, era and composer identity are the same variable, and leave-one-composer-out evaluation is undefined for that class. Because the classes are uneven (Romantic is 63% of clips), all classifiers use `class_weight="balanced"`, all reported metrics are balanced accuracy, and nominal chance is 0.333. Beethoven is labeled Classical throughout unless noted; relabeling him as Romantic is a deliberate control on whether the classifier tracks style or memorizes composers, and both labelings are reported.

3.5 Per-clip feature extraction

Each clip is passed through TRIBE's `predict()`, which returns roughly fifteen retained per-TR predictions per clip; these are mean-pooled over TRs to yield one 20,484-dimensional predicted response vector per clip. The audio baseline runs the identical clips through TRIBE's own Wav2Vec2-BERT extractor, mean-pooled over segments and time, to yield one 1,024-dimensional vector per clip. Because this is the representation TRIBE ingests before its cortical-mapping stage, the paired comparison isolates what the audio→predicted-fMRI mapping adds or removes.

3.6 Decoding protocol

The classifier is a fixed pipeline: per-feature standardization, whitened PCA to n_{pc} components, and multinomial logistic regression (`lbfgs`, `max_iter=2000`, balanced class weights). The primary analyses use $n_{\mathrm{pc}} = 10$, fixed a priori to match the roughly ten-dimensional effective rank of the predicted responses (§3.7); it is a rank-based rule, not a value tuned for accuracy. Because

no accuracy should depend on an unstated component count, n_{pc} is passed explicitly at every call site, and Figure 6 reports the full sweep (5–800) with a convergence audit (Appendix A.6). The localization analysis (§3.8) uses $n_{\mathrm{pc}} = 30$, and the matched-dimensionality controls use $n_{\mathrm{pc}} \in \{5, 30\}$.

We use two cross-validation schemes, and the distinction carries the analysis. *Composers-shared* cross-validation (`StratifiedGroupKFold`, 5 splits, grouped on `composition_id`; shorthand *grouped-CV*) keeps every window of a recording in one fold, but composers are deliberately shared across folds; it is the leaky baseline against which LOCO is contrasted, not a composer-safe split. Leave-one-composer-out (LOCO; 12 folds) holds out every clip by one composer, so era must transfer to a composer the classifier never saw; LOCO is the generalization measure. The grouped-CV−LOCO gap quantifies the composer-identity confound. For the ridge replication (§3.9) we report rank and LOCO only, because its analysis code's "grouped" split is composer-safe and not comparable to grouped-CV here.

3.7 Baselines, permutation nulls, and effective rank

As an external control we compute acoustic descriptors in three blocks (ten timbral/rhythmic descriptors, a 37-dimensional chroma block, and a 26-dimensional MFCC block) on a stratified subset of 1,992 clips, so descriptor numbers are comparable only to one another. These are our own NumPy/SciPy reimplementations, not `librosa` (Appendix A.5).

Two permutation procedures are used. For decodability, we permute era labels (1,000 permutations) under grouped-CV. For cross-composer transfer, we permute the assignment of composers to eras, sampling 500 of the 7,920 distinct composer→era partitions, and recompute LOCO accuracy for each. The *mean* of this null, not 1/3, is the correct reference, because it captures the transfer expected purely from partitioning 12 composers into three unequal groups. We compare the predicted and audio representations against this null and compare the spread of their two nulls with a Pitman–Morgan paired-variance test (Pitman, 1939; Morgan, 1939).

We measure effective dimensionality by singular-value decomposition of the clip-by-feature matrix in two bases: raw (unscaled) and standardized, in which each feature (vertex, voxel, or audio dimension) is z-scored across clips; the standardized basis is what the classifier sees. With $\lambda_1 \geq \lambda_2 \geq \ldots$ the resulting component variances, we report rank@$q = \min\{k : \sum_{i \leq k} \lambda_i \geq q \sum_i \lambda_i\}$ for $q \in \{0.90, 0.95, 0.99\}$ and, as a threshold-free summary, the participation ratio $\mathrm{PR} = (\sum_i \lambda_i)^2 / \sum_i \lambda_i^2$ (Gao et al., 2017). The audio-to-predicted ratio of rank@q quantifies compression. "Effective rank of roughly ten" is shorthand for 9 standardized components at 95% and 10 raw components at 99%. Controls test global redundancy via vertex–vertex correlations across 500 random vertices and whether temporal pooling causes the collapse, via unpooled per-TR predictions for a stratified 240-clip subset (Appendix A.9). Exploratory PC analyses (Appendix A.11) involve informal comparisons without multiplicity correction.

3.8 ROI localization and size-matched random-vertex nulls

To ask whether era decoding localizes to any region of the predicted responses, we decode era under LOCO from each of the 22 coarse cortical divisions of the HCP-MMP1.0 atlas (Glasser et al., 2016) separately (parcels with ≥ 20 vertices), at $n_{\mathrm{pc}} = 30$. A parcel's accuracy is interpretable only relative to what an arbitrary set of the same number of vertices would achieve, so we build two nulls. The *size-matched random-vertex null* decodes era from vertex sets drawn at random from the whole surface, without regard to spatial contiguity, in five size bins spanning the observed parcel range (343–1,653 vertices; 100 draws per bin); it gives a per-size reference for whether a parcel exceeds a random set of equal size. The *max-statistic family-wise null* takes, for each of 100 random parcellations of the surface into 22 parcels, the maximum accuracy over parcels; comparing the observed best parcel with this distribution corrects for having tested all 22. Neither null is designed to reproduce the spatial autocorrelation of the predicted maps (§5.3). Per-parcel accuracies are rendered on the surface for display (Figure 1D).

3.9 Independent-encoder replication: a ridge encoder on ds003020

To test whether the low-rank structure is specific to TRIBE, we fit a second audio-to-fMRI encoder on an unrelated dataset and applied it to the same clips. The encoder is voxelwise ridge regression (`RidgeCV`, $\alpha \in$ `logspace(1,6,11)`, chosen separately for each voxel) trained on ds003020 (LeBel et al., 2023), a public natural-speech fMRI dataset (subject UTS02; 30 training stories and 3 fixed held-out test stories), predicting all 94,251 cortical voxels of that subject. Its input is a single feature space: one 1,024-dimensional mean-pooled Wav2Vec2-BERT vector per non-overlapping 15-second window, the same front end as TRIBE's audio pathway. Because there is one feature space and the penalty is already selected per voxel, banded ridge regression (Dupré la Tour et al., 2022), which assigns separate penalties to distinct feature spaces, reduces to this model. Responses were aligned to the stimulus following HuthLab's pipeline for this dataset (Huth Lab, 2023) (Appendix A.10). The fitted encoder was applied to the cached MAESTRO Wav2Vec2-BERT vectors of §3.5, and its predictions were passed through the same rank and decoding pipeline as TRIBE's.

Two robustness checks accompany the replication. First, we swept the training set from 6 to 30 stories (cumulative; same 3 test stories throughout), tracking held-out encoder quality (voxelwise Pearson r) and prediction rank, to distinguish a genuinely low-rank representation from an under-fit encoder. Second, we tested whether voxels predicted above a shift null (30 circular shifts of the held-out responses; per-voxel 99th percentile) are enriched in the subject's hand-drawn auditory-cortex ROI from ds003020's pycortex database, using a validated reimplementation of pycortex's voxel-to-ROI mapping (Appendix A.10).

4 Results

Sections 4.1–4.2 establish the low-rank structure and its replication, Sections 4.3–4.4 cover era decodability and the absence of localization, and Sections 4.5–4.6 test transfer to held-out composers and compare with acoustic baselines. All accuracies are balanced accuracy (nominal chance 0.333), with $n_{\mathrm{pc}} = 10$ unless stated.

4.1 The predicted responses are effectively low-rank

In the standardized basis the classifier sees, 6, 9, and 19 components explain 90%, 95%, and 99% of the variance of the 20,484-dimensional predicted responses (Figure 1A). In the raw basis the counts are 3, 5, and 10, with 75% of all variance in the first component (participation ratio 1.7). The audio input from which these predictions are computed is markedly higher-dimensional (91/195/529 standardized components; 26/51/213 raw), so the encoding stage compresses its input by 15.2× at 90% variance, 21.7× at 95%, and 27.8× at 99%: a representation of roughly ten effective dimensions emitted into a 20,484-vertex space. This is a measured property of these predictions, not a consequence of any bound relating output rank to input rank. The bottleneck is close to a linear re-encoding of a small audio subspace: the leading predicted components are well predicted from the top 50 audio components in-sample (PC1 R^2 = 0.64, PC3 0.50, PC4 0.47, PC5 0.48).

The redundancy is global rather than pairwise. Over 500 random vertices the median inter-vertex correlation is only +0.33 and just 3.8% of pairs exceed $|r| = 0.9$, so vertices are different linear readouts of the same few latent dimensions rather than duplicates. The collapse is also not an artifact of averaging over time: on unpooled per-TR predictions (Figure 7) the effective rank rises only to 5/7/17 components at 90/95/99% (from 3/4/9 when pooled, same 240 clips), and 65.7% of the TR-level variance is within-clip. Pooling roughly halves the effective rank but does not create the collapse.

4.2 An independently fit encoder is also low-rank

The ridge encoder fit to ds003020 (§3.9) and applied to the same 8,625 clips produces predictions that are lower-rank still: 3, 6, and 16 standardized components explain 90%, 95%, and 99% of their variance, against TRIBE's 6, 9, and 19 (Table 1). Because these predictions are linear in 1,024-dimensional features, their rank cannot exceed 1,024; the observed collapse lies far inside that

Table 1: Dimension, standardized effective rank, and cross-composer transfer for three representations of the same 8,625 clips. LOCO is balanced accuracy (nominal chance 0.333); p is from the 500-partition composer-permutation null (§3.7), which was not computed for the ridge predictions. Grouped-CV is omitted because the two analyses' "grouped" splits differ (§3.6).

Representation	Dim.	rank@90	rank@95	rank@99	LOCO	p
Audio input (Wav2Vec2-BERT)	1,024	91	195	529	0.383	0.30
TRIBE predictions	20,484	6	9	19	0.469	0.004
Ridge predictions (ds003020)	94,251	3	6	16	0.280	–

bound and far below the input's own 91/195/529, so it is produced by the fitted mapping rather than imposed by the bound.

Under LOCO the ridge predictions decode era at 0.280. Without transferable signal, LOCO accuracy is expected to fall somewhat below the nominal 0.333: the composer-permutation nulls of the other two representations are centered at 0.323 and 0.309, with standard deviations of 0.059 and 0.149. We did not compute this null for the ridge predictions, so we read 0.280 as no better than chance rather than as reliably below it. The failure is unsurprising for an encoder fit on spoken narratives and applied out of domain to solo piano, and we do not read it as evidence about linear encoders in general.

Two checks indicate that the ridge encoder's low rank does not reflect a failed fit. First, as its training set grows from 6 to 30 stories, held-out encoder quality rises (95th-percentile voxelwise r: 0.158 $\rightarrow$ 0.171) while prediction rank at 95% variance stays at 6–7 components (Figure 9); rank does not climb as the encoder improves. Second, on the subject's measured held-out responses, the voxels this encoder predicts above a shift null are enriched in auditory cortex: 329 of 2,362 above-null voxels fall in the subject's auditory-cortex ROI against a 5.28% base rate, a 2.64$\times$ enrichment (hypergeometric $p \approx 2.6 \times 10^{-58}$). The encoder captures genuine, if modest, auditory structure.

This comparison is exploratory rather than a controlled ablation. The two encoders differ at once in architecture (deep versus linear), training data (audiovisual movies versus spoken narratives), subjects (group-averaged versus one individual), and output space (fsaverage5 vertices versus native voxels), while sharing the Wav2Vec2-BERT front end, and the ridge is evaluated out of domain. The comparison shows that two very different pipelines both land in a low-rank regime; it does not isolate which factor produces that regime, and the shared front end means it bounds, rather than establishes, independence from the input representation.

4.3 Era is decodable from the predicted responses

Under grouped-CV, a linear classifier recovers the three-way era label from TRIBE's predicted responses with 0.520 balanced accuracy. A 1,000-permutation era-label null is centered at 0.333 $\pm$ 0.008, and no permutation reaches the observed accuracy (permutation p = 0.001, the resolution floor). Era is therefore linearly decodable when whole recordings, but not whole composers, are held out; transfer to held-out composers is tested in §4.5.

4.4 Decoding does not localize: no parcel exceeds a size-matched random-vertex null

Low rank constrains every region to a shared low-dimensional representation but does not by itself imply identical regional performance, so we test whether any parcel exceeds a random vertex set of the same size. None does. Decoding era separately from each of the 22 parcels (Figure 1B,C) yields a narrow band of LOCO accuracies from 0.449 to 0.467, with the whole-cortex value at 0.455. The random-vertex null is flat across a fivefold range of sizes (bin means 0.458–0.461 from 343 to 1,653 vertices). Under the max-statistic null, the best parcel (labeled Early Visual Cortex, 0.467) sits on the null mean of the maximum (0.467 $\pm$ 0.002), giving a family-wise p = 0.465. The result does not hinge on the multiplicity correction: uncorrected, only 1 of 22 parcels exceeds its own size-matched 95th percentile, against 1.1 expected by chance, and the Early Auditory and Auditory Association parcels sit at the 51st and 14th percentiles of their own nulls. For an audio-only input, the top-

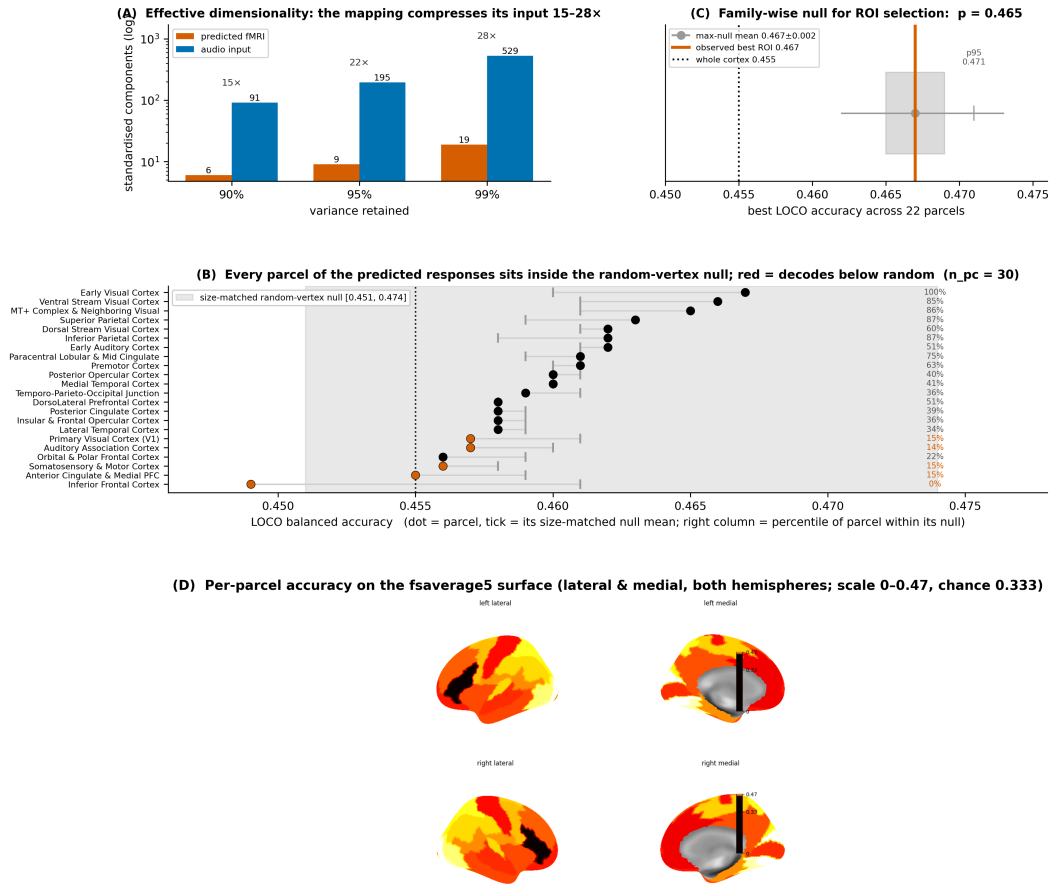


Figure 1: From 20,484 vertices to ∼10 effective dimensions to an apparent, but uninformative, spatial map. **(A)** Standardized components needed to reach 90/95/99% variance for the predicted responses vs the audio input (compression 15–28×). **(B)** Per-parcel LOCO accuracy (n_{pc}= 30); every parcel sits inside the size-matched random-vertex null band, red marks parcels decoding below random, and the right column gives each parcel's percentile within its own null. **(C)** Family-wise max-statistic null: the best parcel (0.467) lands on the null mean, family-wise $p = 0.465$. **(D)** The same per-parcel accuracies on the fsaverage5 surface (lateral and medial views; color scale 0–0.47, chance 0.333). The map is smooth and reads like a localization result, yet no parcel exceeds the size-matched random-vertex null.

ranked parcel is a visual one. Rendered on the surface (Figure 1D), however, these accuracies form a smooth, plausibly organized map.

4.5 Secondary: Transfer to Held-Out Composers

Holding out whole composers lowers predicted-response accuracy from 0.520 to 0.469 and audio accuracy from 0.686 to 0.383 (Figure 2A). The mean difference of +0.085 in favor of the predicted responses is not distinguishable from zero under composer-to-era permutation (p = 0.36), and its sign reverses at $n_{\text{pc}} = 800$. The two representations differ instead in how consistently era transfers. Against the composer-partition null, transfer is significant for the predicted responses (0.469 vs 0.323 $\pm$ 0.059, p = 0.004) but not for audio (0.383 vs 0.309 $\pm$ 0.149, p = 0.30). Per held-out composer, the fraction of clips classified correctly ranges from 0.208 to 0.719 for the predicted responses (SD 0.139) but from 0.000 to 1.000 for audio (SD 0.343; Figure 2B), and the audio null is about 2.5 times as dispersed (Pitman–Morgan test, Appendix B.2). This is a difference

in the consistency of transfer across composers, not evidence that the predictions carry more era information.

4.6 Secondary: no advantage over acoustic baselines

On the 1,992-clip subset (§3.7), we compared the predicted responses under LOCO with acoustic descriptor blocks (Figure 4). At matched dimensionality, acoustic baselines equal or exceed the predicted responses at every setting: the predicted responses reach 0.420 at five components and 0.476 at thirty, versus 0.483 and 0.508 for the combined acoustic block; chroma-plus-MFCC reaches 0.493 at five components. A five-component projection of the predicted responses (0.428) also scores no higher than the full 20,484-vertex representation. There is thus no evidence that the predictions preserve era information beyond standard acoustic features; the point estimates favor the acoustic baselines, but the gaps are small and generalization is limited by the twelve-composer sample.

5 Discussion

5.1 Summary and recommendation

Era was decodable from TRIBE's predicted responses, and its transfer to held-out composers exceeded a composer-permutation null, whereas the input audio representation's transfer did not and varied far more across composers; the mean accuracies did not differ significantly. The predictions nonetheless occupied a highly constrained subspace, with an effective rank of roughly ten in a 20,484-vertex space. An independently fit ridge encoder, differing in architecture, training data, and subject, was lower-rank still. Parcel-level decoding showed no evidence of regional specialization relative to size-matched random-vertex nulls, and the best parcel did not exceed the family-wise max-statistic null ($p = 0.465$). Spatially organized decoding maps can thus arise from a low-dimensional representation without evidence for regionally specific information. We recommend two diagnostics for downstream analyses of predicted responses: assess effective rank in the standardized representation the analysis uses, and evaluate regional effects against random vertex sets matched for size.

5.2 Why low rank limits localization

Write the $n \times p$ matrix of predicted responses as $\hat{Y} = ZB^\top + E$, where $Z \in \mathbb{R}^{n \times k}$ holds the scores of $k \ll p$ latent components, $B \in \mathbb{R}^{p \times k}$ their vertex loadings, and E a low-variance residual. For any vertex set S, $\hat{Y}_S = ZB_S^\top + E_S$. Whenever B_S has full column rank, which holds generically once $|S| \geq k$, the latents are recovered linearly: $(\hat{Y}_S - E_S)\, B_S (B_S^\top B_S)^{-1} = Z$. Every such set, contiguous or scattered, therefore gives a linear decoder access to the same latent variables. Sets can differ only in how B_S weights those latents, which shapes variance-based preprocessing such as PCA truncation and whitening, and in their residual E_S. Parcels here contain 343–1,653 vertices, one to two orders of magnitude more than the 19 standardized components that carry 99% of the variance, and the observed parcel accuracies span only 0.449–0.467. Low rank does not prevent decoding: separating three classes linearly requires at most a two-dimensional discriminant subspace, which a rank-$\sim$10 representation easily contains. What low rank constrains is region-specific information: with roughly ten latents shared by 20,484 vertices, a parcel of a few hundred vertices generically reads out all of them, so parcels can differ only through weighting and residual. This is how predictions can support decoding without supporting localization.

5.3 What the random-vertex null does and does not test

The random-vertex null draws vertices without regard to spatial contiguity, so it is an aspatial, size-matched baseline and not a spatial-autocorrelation-preserving null of the kind used to compare brain maps (Alexander-Bloch et al., 2018; Burt et al., 2020; Markello & Mišić, 2021; Markello et al., 2022). It tests whether a parcel carries more era information than an arbitrary vertex set of the same size, which is the comparison a localization claim implicitly makes. A scattered set could in principle sample more latent dimensions than a contiguous parcel whose smooth loadings make B_S ill-conditioned, making the test conservative toward parcels. Two observations limit this concern here. First, spatial dispersion confers no measurable advantage: the whole cortex, which contains every

dimension any subset could sample, decodes no better (0.455) than random subsets (0.458–0.461) or the parcels themselves, and random-set accuracy is flat across a fivefold range of sizes. Second, the non-result does not depend on the family-wise correction (§4.4). A contiguity-preserving null, such as spin-rotated parcels or size-matched geodesic patches, would ask a complementary question, namely whether a parcel outperforms other contiguous regions of similar shape, and we did not run one. Our claim is accordingly that no parcel carries more era information than an arbitrary size-matched vertex set, not that every spatial null would be passed. When a spatial claim is at stake, we recommend reporting per-parcel effective rank alongside both kinds of null.

5.4 Limitations

No measured baseline. The piano clips were never scanned, so we cannot say whether an effective rank of roughly ten is lower than that of measured whole-cortex responses to the same music. Measured responses are themselves low-dimensional in some respects (Norman-Haignere et al., 2015), and era may be encoded in a distributed rather than localized way in real cortex. Our argument does not depend on either question: whenever a representation is low-rank relative to its vertex count, regional comparisons on it require such a null. Applying the same rank and parcel-versus-null pipeline to measured responses to music, for example the fifteen-second clips of ds003720 (Nakai et al., 2022), is the natural next step for separating model-induced from biological structure.

Out-of-domain use. TRIBE was trained on audiovisual movies and is driven here by audio alone through its modality-dropout pathway, and the ridge encoder was trained on speech. Both are therefore applied outside their training distribution. That is the setting of the practice we examine, namely predicting responses to stimuli that were never scanned, but the magnitude of the collapse may depend on it; we did not measure either encoder's prediction rank on in-distribution held-out stimuli.

Design. The primary analyses use temporally pooled predictions; the unpooled control shows that pooling reduces rank without creating the collapse, but TR-level decoding remains untested. Generalization is evaluated over 12 composers rather than 8,625 clips, so composer-level nulls are used and differences unresolved at that level are not interpreted. Beethoven's era label is contestable, and both labelings are retained. Chronology and single-PC analyses are exploratory and do not inform the localization claim. Appendix C expands on these points.

5.5 Scope and conclusion

In this experiment, a 20,484-vertex predicted response occupied an effective subspace of roughly ten standardized dimensions, an independently fit encoder on an unrelated dataset and subject was lower-rank still, and no cortical parcel exceeded a size-matched random-vertex null. Anatomical indexing of predicted responses therefore does not by itself establish that their decodable information is spatially localized or regionally specialized. These findings do not bear on the validity of encoding models, do not imply that measured fMRI has the same rank structure, and provide no evidence about human cognition. They identify an interpretive limitation that arises when predicted responses are analyzed according to their anatomical layout, and effective-rank reporting together with size-matched nulls offer an inexpensive safeguard.

Reproducibility statement

Section 3 and Appendix A specify the audio-only inference path (§3.3), stimulus construction (§3.4), the classifier and its a-priori component count (§3.6), both cross-validation schemes, the permutation nulls and the rank definitions (§3.7), the ROI procedure and its two nulls (§3.8), and the ridge replication with its robustness checks (§3.9, Appendix A.10). Seeds, software versions, and compute are listed in Appendix A.12. The analysis notebooks (as `.py`), data-preparation and feature-extraction scripts, figures, cell outputs, and permutation-null arrays are available in an anonymized supplementary repository linked from the submission site; code will be released publicly upon acceptance. MAESTRO (CC BY-NC-SA 4.0), TRIBE v2 (CC BY-NC 4.0), and ds003020 (LeBel et al., 2023) are public and are not redistributed; any released features derived from MAESTRO inherit its license. No lyrics or musical scores are reproduced.

Ethics statement

This work involves no new data collection, no human participants, and no personally identifying information. Every quantity reported for the piano stimuli is a deterministic function of publicly licensed audio (MAESTRO) passed through public or openly trained encoding models. The ridge replication fits an encoder to a public, previously collected, de-identified natural-speech fMRI dataset, ds003020 (LeBel et al., 2023), used only to obtain a second encoder and not to make claims about that dataset or its participants. No institutional review was required (§3.1).

AI use statement

We used LLM-based assistants in three ways. *Research ideation and execution:* assistants suggested analyses and controls and helped write and debug analysis code, including the ds003020 ridge-encoder replication and its robustness checks (§3.9); they also helped formulate the linear-algebra argument of §5.2 and the interpretation of the random-vertex null in §5.3. *Drafting:* assistants drafted parts of the text from our results, including the ridge-replication methods and results, §5.2–§5.3, and the limitations. *Writing polish:* assistants edited for clarity, converted the manuscript to the ICLR template, formatted references, and checked anonymization. LLMs were not used to generate data or to produce or alter any reported number; every number is the output of executed analysis code. We reviewed all AI-assisted code, derivations, and text, and every added reference was checked against its publisher record. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

References

Aaron F. Alexander-Bloch, Haochang Shou, Siyuan Liu, Theodore D. Satterthwaite, David C. Glahn, Russell T. Shinohara, Simon N. Vandekar, and Armin Raznahan. On testing for spatial correspondence between maps of human brain structure and function. *NeuroImage*, 178:540–551, 2018. doi: 10.1016/j.neuroimage.2018.05.070.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. In *Interspeech*, 2023.

Loïc Barrault et al. Seamless: Multilingual Expressive and Streaming Speech Translation. *arXiv preprint arXiv:2312.05187*, 2023. Wav2Vec2-BERT speech encoder.

Joshua B. Burt, Markus Helmer, Maxwell Shinn, Alan Anticevic, and John D. Murray. Generative modeling of brain maps with spatial autocorrelation. *NeuroImage*, 220:117038, 2020. doi: 10.1016/j.neuroimage.2020.117038.

Stéphane d'Ascoli, Jérémy Rapin, Yohann Benchetrit, Teon Brooks, Katelyn Begany, Joséphine Raugel, Hubert Banville, and Jean-Rémi King. A foundation model of vision, audition, and language for in-silico neuroscience. *arXiv preprint arXiv:2605.04326*, 2026. TRIBE v2.

Tom Dupré la Tour, Michael Eickenberg, Anwar O. Nunez-Elizalde, and Jack L. Gallant. Feature-space selection with banded ridge regression. *NeuroImage*, 264:119728, 2022. doi: 10.1016/j.neuroimage.2022.119728.

Peiran Gao, Eric Trautmann, Byron M. Yu, Gopal Santhanam, Stephen Ryu, Krishna Shenoy, and Surya Ganguli. A theory of multineuronal dimensionality, dynamics and measurement. *bioRxiv preprint 214262*, 2017. doi: 10.1101/214262.

Alessandro T. Gifford, Domenic Bersch, Marie St-Laurent, Basile Pinsard, Julie Boyle, Lune Bellec, Aude Oliva, Gemma Roig, and Radoslaw M. Cichy. The Algonauts Project 2025 Challenge: How the Human Brain Makes Sense of Multimodal Movies. *arXiv preprint arXiv:2501.00504*, 2025.

Matthew F. Glasser, Timothy S. Coalson, Emma C. Robinson, Carl D. Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F. Beckmann, Mark Jenkinson, Stephen M. Smith, and David C. Van Essen. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016. doi: 10.1038/nature18933.

Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset. In *International Conference on Learning Representations (ICLR)*, 2019.

Huth Lab. deep-fMRI-dataset: code accompanying the natural-language-listening fMRI data release (LeBel et al.). https://github.com/HuthLab/deep-fMRI-dataset, 2023.

Amanda LeBel, Lauren Wagner, Shailee Jain, Aneesh Adhikari-Desai, Bhavin Gupta, Allyson Morgenthal, Jerry Tang, Lixiang Xu, and Alexander G. Huth. A natural language fMRI dataset for voxelwise encoding models. *Scientific Data*, 10:555, 2023. doi: 10.1038/s41597-023-02437-z.

Ross D. Markello and Bratislav Mišić. Comparing spatial null models for brain maps. *NeuroImage*, 236:118052, 2021. doi: 10.1016/j.neuroimage.2021.118052.

Ross D. Markello, Justine Y. Hansen, Zhen-Qi Liu, Vincent Bazinet, Golia Shafiei, Laura E. Suárez, Nadia Blostein, Jakob Seidlitz, Sylvain Baillet, Theodore D. Satterthwaite, M. Mallar Chakravarty, Armin Raznahan, and Bratislav Mišić. neuromaps: structural and functional interpretation of brain maps. *Nature Methods*, 19(11):1472–1479, 2022. doi: 10.1038/s41592-022-01625-w.

William A. Morgan. A test for the significance of the difference between the two variances in a sample from a normal bivariate population. *Biometrika*, 31(1/2):13–19, 1939.

Tomoya Nakai, Naoko Koide-Majima, and Shinji Nishimoto. Music genre neuroimaging dataset. *Data in Brief*, 40:107675, 2022. doi: 10.1016/j.dib.2021.107675.

Thomas Naselaris, Kendrick N. Kay, Shinji Nishimoto, and Jack L. Gallant. Encoding and decoding in fMRI. *NeuroImage*, 56(2):400–410, 2011. doi: 10.1016/j.neuroimage.2010.07.073.

Sam Norman-Haignere, Nancy G. Kanwisher, and Josh H. McDermott. Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *Neuron*, 88(6):1281–1296, 2015. doi: 10.1016/j.neuron.2015.11.035.

Edwin J. G. Pitman. A note on normal correlation. *Biometrika*, 31(1/2):9–12, 1939.

Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Matteo Carandini, and Kenneth D. Harris. High-dimensional geometry of population responses in visual cortex. *Nature*, 571(7765):361–365, 2019. doi: 10.1038/s41586-019-1346-5.

George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002. doi: 10.1109/TSA.2002.800560.

A Supplementary methods

A.1 Encoding model

TRIBE's configured fMRI sampling frequency is 0.671 Hz (TR $\approx$ 1.49 s), and the audio pathway's Wav2Vec2-BERT feature tensor has shape `(n_segments, 1024, T)`. Figure 3 shows the full audio-to-decoding pipeline.

A.2 Audio-only inference

This section expands §3.3. The convenience entry point is `model.get_events_dataframe(audio_path=...)`, which internally calls `get_audio_and_text_events()` *without* `audio_only=True` and therefore runs the WhisperX (Bain et al., 2023) large-v3 ASR stage. On instrumental piano that stage hallucinates lyrics, and the spurious `Word` events then feed the model's text branch and contaminate its predictions. Calling the module-level function directly with `audio_only=True` skips ASR entirely; with no `Word` events present, the loader drops the text and video extractors and `FmriEncoder.aggregate_features` substitutes zero tensors for the absent modalities. We confirmed at runtime that the text and video extractors are removed and only the audio extractor is prepared, in addition to a per-batch assertion that no `Word` events are present.

A.3 Stimuli

A `DROP_SINGLE_COMPOSER_ERAS` rule enforces that every retained era contains at least two composers, which is what removes Debussy as the sole Impressionist/Modern composer. The Beethoven relabeling control is implemented by the `BEETHOVEN_AS_ROMANTIC` setting. There are no human-subjects data and no IRB requirement.

A.4 Per-clip feature extraction

About 15% of 100 candidate segments survive `remove_empty_segments`, giving the roughly fifteen retained per-TR predictions per clip. Segment-to-clip demultiplexing uses `segment.ns_events[0].timeline`, with each clip's timeline set to its `window_uid`, the same mechanism TRIBE uses internally. The audio baseline uses `model.data.audio_feature`.

A.5 Acoustic descriptors

The descriptor formulations are standard and adequate for correlation and classification, but they are not numerically identical to librosa's, and all descriptor analyses were run on a stratified subset of 1,992 clips (targeting ~2,000), so descriptor numbers are comparable only to one another, not to the full-data figures. Three feature blocks are used: ten timbral/rhythmic descriptors (spectral centroid, bandwidth, roll-off, flatness, RMS energy, zero-crossing rate, onset rate, tempo, harmonic ratio, and dynamic range); a chroma block (12 chroma means, 12 key-invariant/rotated chroma, 12 chroma standard deviations, and chroma entropy, totalling 37); and an MFCC block (13 MFCC means and 13 MFCC standard deviations, totalling 26). Harmonic–percussive separation is a simplified median-filter variant, and tempo is estimated by autocorrelation of the spectral-flux onset envelope rather than by a dedicated beat tracker. `librosa` was unusable in the environment because its `numba` dependency requires NumPy $\leq$ 2.0 and the environment ran NumPy 2.2.

A.6 Decoding protocol: provenance and component-count auditing

The classifier uses `StandardScaler`, `PCA(whiten=True)`, and `LogisticRegression(solver="lbfgs", max_iter=2000, class_weight="balanced")`; LOCO uses `LeaveOneGroupOut` on composer. Grouped-CV (composers-shared, §3.6) uses `shuffle=True` and `random_state=0`. During analysis we discovered that the classifier factory's PCA component count had drifted from its intended value, which had moved reported accuracies without warning. In the final code n_{pc} is an explicit,

documented argument guarded by a provenance-check cell and checkpoint-metadata assertions, so every reported accuracy is tied to a stated component count, and the choice is auditable rather than load-bearing. Figure 6 also audits convergence, since lbfgs can exhaust `max_iter` once n_{pc} exceeds the effective rank; all columns converged (max `n_iter` ≤ 72). Neither cross-validation scheme's accuracy plateaus over the 5–800 range, which is why the curve rather than a single point is reported.

A.7 Permutation-test efficiency

For efficiency the standardizer and PCA are fit once on the full matrix and only the classifier is refit per permutation; both transforms are unsupervised and applied identically to real and permuted labels, and we verified that this shortcut reproduces the non-permuted LOCO accuracy to $|\Delta| \leq 0.001$ for both representations.

A.8 Effective-rank analysis

Rank@q and the participation ratio are defined in §3.7; both are computed from the singular values of the clip-by-feature matrix, so no choice of component count enters them. To test whether the predicted-response bottleneck is a linear re-encoding of a small audio subspace, we regress each leading predicted-response principal component on the top-50 audio principal components and report in-sample R^2.

A.9 Unpooled (TR-level) rank control

Mean-pooling ~15 per-TR predictions per clip removes temporal variance and could by itself lower the effective rank, so we ran a control on the unpooled predictions. On a stratified 240-clip subset we repeated the exact audio-only forward pass of §3.3 but retained the individual per-TR predictions rather than their per-clip mean, comparing the effective rank of unpooled and pooled predictions on the same clips and decomposing the TR-level variance into between-clip and within-clip components. This control requires a fresh GPU forward pass and shares the call path and the no-`Word` assertion of §3.3.

A.10 Independent ridge encoder: alignment and ROI mapping

Alignment. Response alignment follows HuthLab's encoding pipeline (Huth Lab, 2023) for ds003020: response-array row j corresponds to stimulus playback time $2j + 11.0$ s (TR = 2.0 s). The offset is derived from `TRFile.soundstarttime`, the Lanczos interpolation center, and the default `-trim 5` feature-grid crop, and was verified empirically at zero mismatch across all 21 available stories. The MAESTRO-side predictions reuse the cached Wav2Vec2-BERT vectors of §3.5; because the original MAESTRO clip audio no longer exists, we could not confirm that the reference `facebook/w2v-bert-2.0` extractor used for ds003020 matches the MAESTRO-side extractor exactly.

ROI mapping. The compiled `pycortex` package would not build in our environment (its `openctm` C extension requires an unavailable Windows toolchain), so we reimplemented its documented voxel-to-ROI mapping from source (`cortex/utils.py`: nearest-neighbor assignment from each functional voxel to the fiducial, i.e. (pia+wm)/2, surface via a k-d tree, restricted to the `mask_thick` cortical mask). The reimplementation reproduces the expected voxel count (94,251, matching the response array) with a 2.24 mm median voxel-to-surface assignment distance. Of the 94,251 voxels, 4,972 (5.28%) fall in the auditory-cortex ROI; 2,362 (2.51%) exceed the shift null, 329 of them in auditory cortex. We rely on the enrichment statistic, not on any rendered flatmap, as evidence.

A.11 Exploratory principal-component and chronology analyses

Two further analyses are reported as exploratory. To relate the predicted-response principal components to composer chronology, we correlate composer-mean PC scores with composer birth year (Spearman ρ). To estimate how much era information a single PC carries without post-hoc component selection, we run a nested single-PC LOCO in which the PC is chosen within each training fold;

this nested estimate, rather than the best PC chosen on the full data, is the selection-honest figure. These analyses, along with the per-descriptor and per-feature-set comparisons, involve many informal comparisons; we apply no formal multiplicity correction and treat them as exploratory rather than confirmatory.

A.12 Reproducibility, software, and data

Random seeds are fixed throughout: subsampling and permutations use `numpy.random.default_rng(0)`, `StratifiedGroupKFold` uses `random_state=0`, and PCA uses `random_state=0`. The predicted-response feature bundle (`clip_fmri_features.npz`) stores the 8,625 × 20,484 matrix together with era labels, `composition_id` groups, and `window_uids`; audio and acoustic features are cached separately. Software versions are as follows: Python 3.12.13, scikit-learn 1.6.1, NumPy 2.2.6, SciPy 1.16.3, nilearn 0.14.0, PyTorch 2.11.0, and TRIBE v2 (`tribev2` 0.1.0). Computations were run on a single L4 GPU. Two environment constraints should be considered for reproduction: TRIBE has no XLA path, so TPU runtimes fall back to CPU and fail; and the audio encoder must be prepared once and reused rather than reloaded per call, which is roughly a 100× speed-up.

B Supplementary results

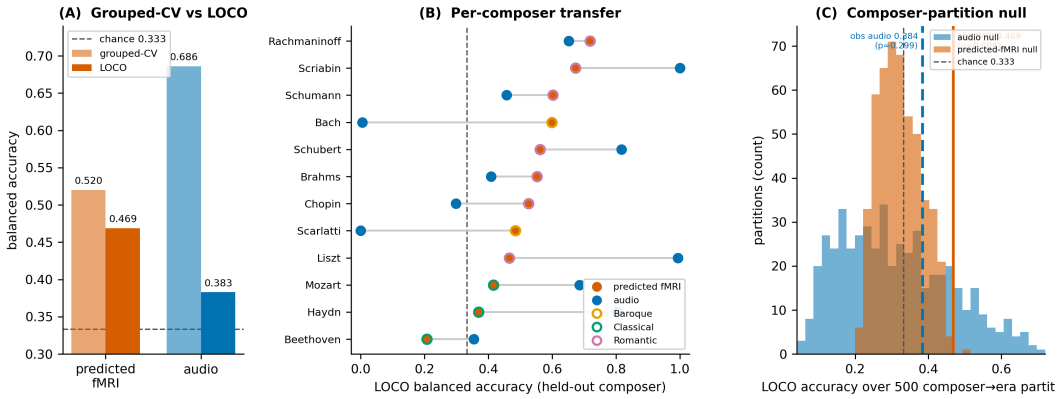


Figure 2: Era transfer to held-out composers ($n_{\mathrm{pc}}= 10$). **(A)** Grouped-CV (composers-shared) vs leave-one-composer-out balanced accuracy for the predicted-response and audio representations (grouped−LOCO gap +0.051 vs +0.303). **(B)** Per-composer LOCO accuracy; each grey line joins a composer's audio and predicted-response value, ring color marks era. **(C)** Composer-partition permutation null (500 partitions) as overlaid histograms for the two representations; vertical lines mark the observed values and chance.

B.1 Per-composer transfer: a graceful degradation against a Baroque collapse

The per-composer breakdown (Figure 2B) shows the same contrast at the level of individual held-out composers. The audio representation's Baroque class collapses entirely under LOCO, with Scarlatti at 0.0% and Bach at 0.5% correct, while the predicted responses hold those same composers at 48.5% and 59.9%. Conversely, the audio representation is near-perfect on several held-out Romantic and Classical composers (Scriabin 100%, Liszt 99.4%, Schubert 81.7%, Haydn 80.3%) where the predicted responses are merely good. The predicted responses degrade gracefully across all twelve composers (range 20.8%–71.9%, SD 0.139), whereas the audio swings the full width of the scale (range 0.0%–100%, SD 0.343). The tails of the two composer-partition nulls make the same point: 7.4% of partitions score above 0.55 for the audio representation against 0.0% for the predicted responses, and the audio null reaches 0.708 where the predicted-response null tops out at 0.508. This is the double picture behind the summary statistics of §4.5: comparable central tendency, very different consistency across composers.

B.2 Summary conventions and tests of central tendency

Averaged over the three eras, the predicted responses lead (0.469 vs 0.383); averaged over the twelve composers, the audio leads (0.540 vs 0.515), and by every test of central tendency the two are indistinguishable at this sample size (composer-macro difference -0.025, 95% CI $[-0.150, +0.091]$; Wilcoxon $W = 33$, $p = 0.68$; sign test 6/12, $p = 1.00$). The two averages disagree because the audio representation's collapse falls almost entirely on the Baroque era, which is one of three eras but only two of twelve composers, while seven of twelve composers are Romantic, where audio is strongest. The Pitman–Morgan paired-variance test on the two 500-partition nulls confirms the audio null is the more variable ($r = 0.747$, $p < 0.001$); in the main text we report the standard-deviation ratio rather than this p-value, which is computed on permutation draws whose number we chose and is not an independent measure of effect size.

B.3 Interpretability vignette: Beethoven

One exploratory observation recurs (Figure 5) across three independent analyses of the predicted representation. First, Beethoven is the worst-transferring composer under LOCO (20.8% correct), and the representation's modal prediction for his clips is Romantic rather than his Classical label. Second, in a nested single-PC analysis, where the most era-informative component is chosen within each training fold and the selection-honest LOCO accuracy is 0.439 with the fifth component chosen in 11 of 12 folds, Beethoven is again the lowest at 14.5%. Third, the leading predicted-response components order the twelve composers by birth year (Spearman $\rho = -0.73$ on PC1, $p = 0.007$; -0.71 on PC5, $p = 0.009$; -0.67 on PC3, $p = 0.018$), and on that chronological axis Beethoven (born 1770) is displaced roughly forty to sixty years forward, sitting among composers born in the 1810s–1830s. The convergence of three separate readouts on the same composer is the strongest interpretability signal in the project, and it is exploratory: it is a property of where this encoding model's predictions place Beethoven, an observation about the model's outputs, and it licenses no claim about music history, stylistic influence, or how the composer is heard.

C Extended limitations

This section expands §5.4.

Measured baseline. No measured responses to the MAESTRO clips exist, so the effective rank of the predictions cannot be compared with that of measured responses to identical stimuli. Measured auditory-cortex responses to natural sounds are dominated by a handful of components (Norman-Haignere et al., 2015), so a low rank would not by itself distinguish predictions from measurements; our claim concerns what low rank implies for regional comparisons, whatever its origin. Running the rank analysis and the parcel-versus-null pipeline on a measured music dataset such as ds003720 (Nakai et al., 2022), whose stimuli are also fifteen-second clips, would show whether measured responses retain regional specificity where predictions do not, or whether era information is distributed in measured cortex as well.

Spatial nulls. The random-vertex and random-parcellation nulls do not reproduce the spatial autocorrelation of the predicted maps (§5.3). Spin-permuted parcels (Alexander-Bloch et al., 2018), variogram-matched surrogates (Burt et al., 2020), or size-matched contiguous geodesic patches would complement them, as would per-parcel effective rank, which identifies parcels whose smooth loadings span fewer latent dimensions than a scattered vertex set of equal size.

Out-of-distribution usage. Measuring effective rank under full multimodal input, and on in-distribution held-out stimuli for both encoders, would separate the contribution of the single-modality, out-of-domain regime from that of the encoding stage itself.

Temporal pooling. All decoding reported here is on pooled features; TR-level decoding is left to future work.

Implementation notes. The acoustic descriptors are our own NumPy/SciPy reimplementations, adequate standard formulations but not numerically identical to librosa, which should not be cited for them. The permutation tests fit the standardizer and PCA once on the full matrix and refit only the classifier per permutation; this is valid because both transforms are unsupervised and applied iden-

tically to real and permuted labels, and we verified that it reproduces the non-permuted accuracy to $|\Delta| \leq 0.001$.

Confounds and analytic choices. Relabeling Beethoven as Romantic shifts the class prior, mechanically moving even Baroque composers whose data are unchanged, which is why both labelings are reported rather than the higher-scoring one. In the exploratory analyses the era-informative component was selected post hoc, so we report the selection-honest nested estimate, and across the many informal comparisons (several component-by-descriptor and component-by-chronology contrasts, six feature sets) we applied no formal multiplicity correction and label these analyses exploratory. The 22-parcel localization analysis is the exception: it is corrected by the max-statistic family-wise null.

Independent-encoder replication. The ridge replication varies architecture, training data, subject, and output space simultaneously, without single-factor ablations, and shares the Wav2Vec2-BERT feature extractor with TRIBE; it therefore bounds rather than establishes generality across encoders independent of the input representation. A linear encoder fit on the same features and training data as TRIBE, or both encoders evaluated on a shared in-distribution test set, would isolate the architectural contribution. The ridge encoder uses a single feature space with per-voxel penalties, so banded ridge regression (Dupré la Tour et al., 2022) would coincide with it; a multi-feature-space variant (for example adding spectrotemporal or music-specific features) is a natural extension. The match between the ds003020-side and MAESTRO-side Wav2Vec2-BERT extraction is unverified, and the auditory-cortex enrichment check relies on a reimplementation of pycortex's voxel-to-ROI mapping, validated on voxel count and median assignment distance.

D Broader Impacts

The work is methodological. Its intended effect is to make localization and region-comparison claims computed on encoding-model predictions harder to make carelessly, by supplying a cheap check (report the effective rank and a size-matched random-vertex null) that costs seconds and can be adopted immediately. The main risk we see is over-generalization in the other direction, and §5.5 states those boundaries. No human-subjects data are involved and no personally identifying information is used or released.

E Supplementary Figures

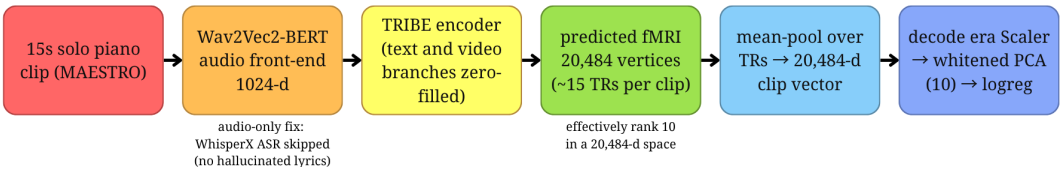


Figure 3: Audio-to-decoding pipeline. Fifteen-second solo-piano clips pass through TRIBE with the audio-only fix enabled (§3.3): the WhisperX ASR stage is skipped and the text and video branches are zero-filled, so the model emits a 20,484-vertex predicted response per TR, which is mean-pooled over the ∼15 retained TRs into one clip vector and decoded. (Encoding model: TRIBE v2.)

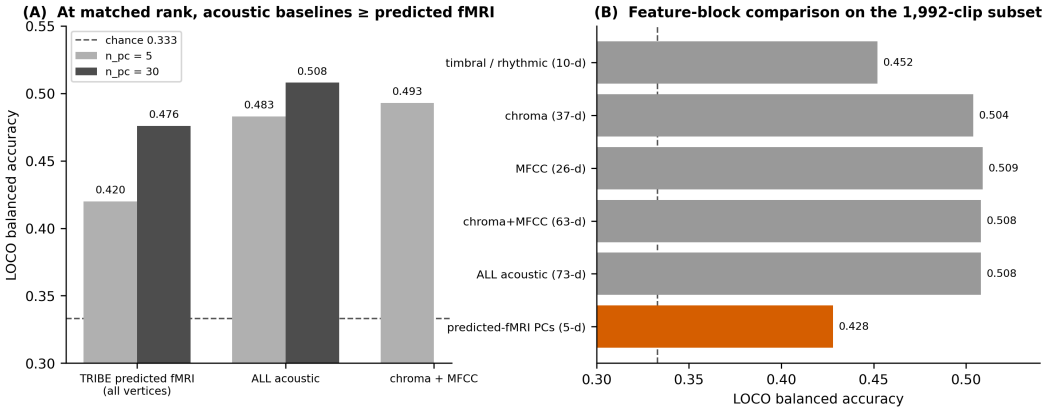


Figure 4: Predicted responses do not decode era beyond acoustic baselines. **(A)** Matched-dimensionality LOCO accuracy at $n_{\mathrm{pc}} \in \{5, 30\}$. **(B)** Per-feature-block LOCO accuracy on the 1,992-clip subset; the classifier reduces each block to at most 30 PCs, but the TRIBE-fMRI representation enters through the 5-PC bottleneck used throughout the acoustic-control analyses, so its point is a 5-PC readout while the acoustic blocks use up to 30. Point estimates favor the acoustic features; gaps are $\sim$0.02–0.05 with an effective $n = 12$ composers.

Figure 5. Three independent readouts of the predictions converge on Beethoven (worst LOCO transfer, worst nested single-PC, displaced forward in chronology)

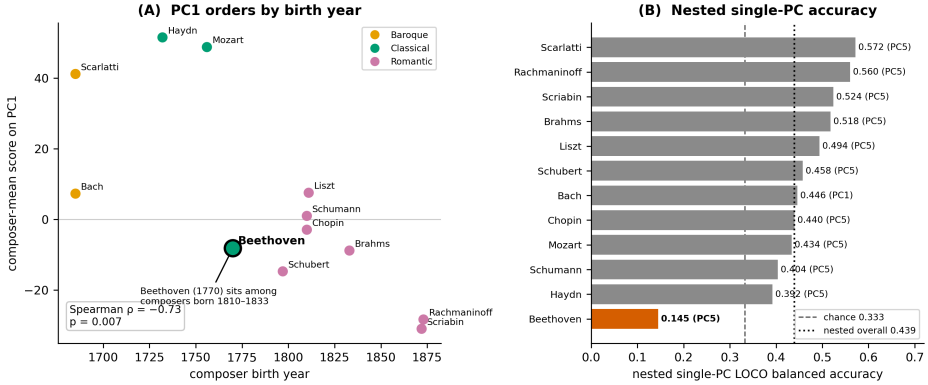


Figure 5: Three independent readouts of the predictions converge on Beethoven. **(A)** Composer-mean score on PC1 vs birth year (Spearman $\rho = -0.73$); Beethoven is displaced forward. **(B)** Nested single-PC LOCO accuracy per composer; Beethoven is lowest. This describes where the model's predictions place one composer and licenses no claim about music history.

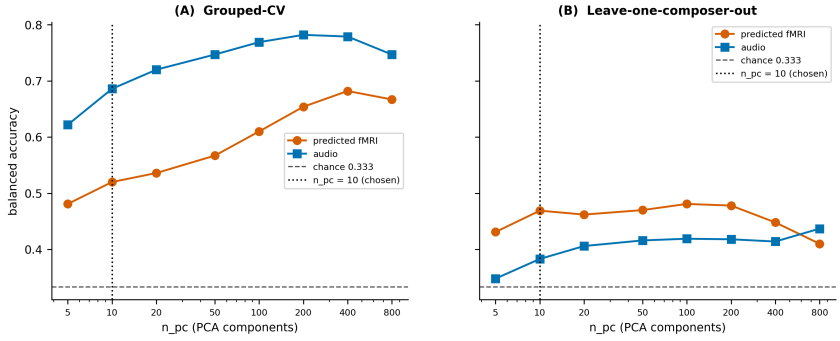


Figure 6: Balanced accuracy vs retained dimensionality n_{pc} for grouped-CV (composers-shared; A) and LOCO (B). Neither scheme plateaus, motivating reporting the curve rather than a single point; the chosen $n_{\mathrm{pc}} = 10$ is marked. All columns converged (max `n_iter`$\leq$72).

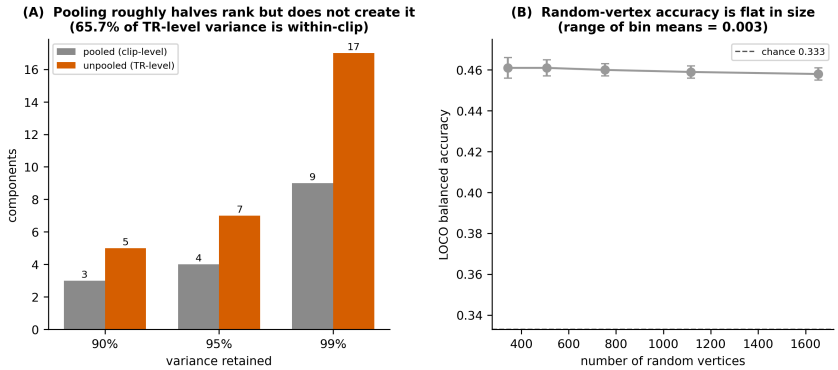


Figure 7: **(A)** Effective rank of unpooled (TR-level) vs pooled predictions: pooling roughly halves rank but does not create the collapse (65.7% of TR-level variance is within-clip). **(B)** Random-vertex accuracy is flat in region size (range of bin means $= 0.003$).

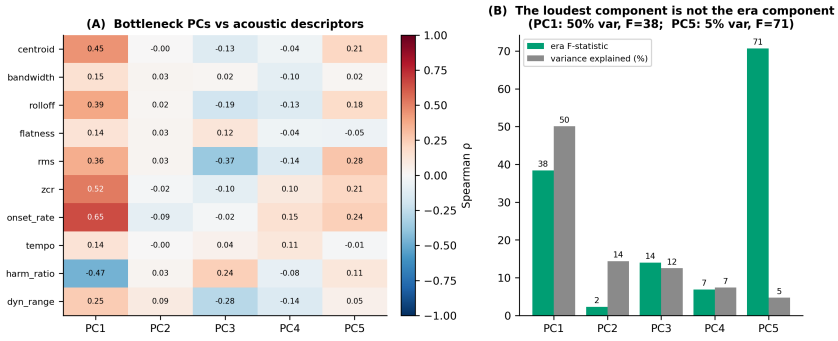


Figure 8: **(A)** Spearman correlations between the leading predicted-response PCs and acoustic descriptors. **(B)** Per-PC era F-statistic vs variance explained: the highest-variance component (PC1) is not the most era-discriminative (PC5).

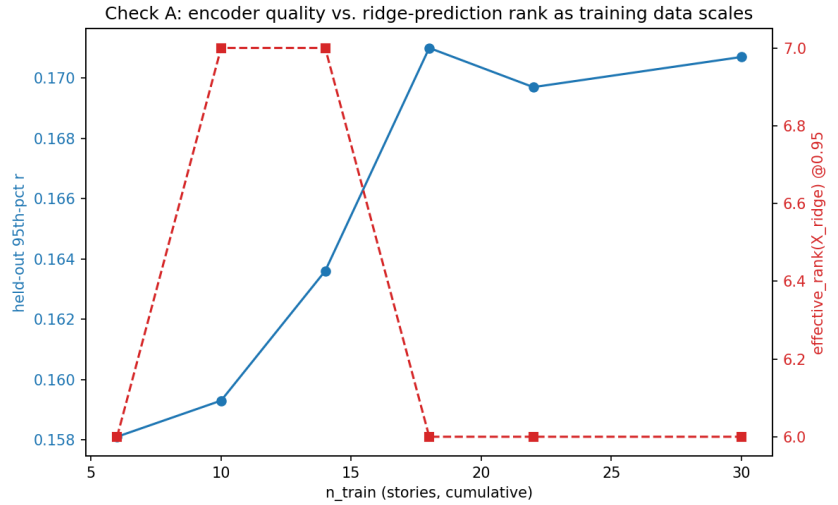


Figure 9: Independent-encoder replication, robustness check (§3.9, §4.2): as the ridge encoder's training set grows from 6 to 30 ds003020 stories, held-out encoder quality (95th-percentile voxel-wise r, left axis) rises from 0.158 to 0.171 while its prediction rank at 95% variance (right axis) stays at 6–7 components throughout. Rank does not climb as the encoder improves, indicating the low-rank collapse is not an artifact of an under-fit encoder.