ECHO: EMBEDDING CONVERGENCE AND HIDDEN ORGANIZATION IN AUDIO MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Pretrained models trained once on large-scale datasets are now routinely repurposed for entirely new tasks, but there is no reliable way to predict, in advance, how much a given model will benefit from adaptation without first fine-tuning it. This is a real cost in domains where labeled data is scarce, forcing researchers to spend weeks and compute on trial-and-error. We propose ECHO: a concrete, falsifiable geometric signature of adaptation headroom, together with a precise map of the tasks where that signature holds and where it does not. We show that a single measurable property of a model's internal representation; independent of its training data, architecture, or task performance predicts how much a model stands to gain from adapting to a new task, before any adaptation is performed. We study numerous independently trained audio models spanning speech, music, environmental sound, bioacoustics, and machine acoustics. We measure how each model internally arranges sound relative to every other model, and find that agreement between models is real but only partially explained by what they were trained on, no single factor accounts for it fully. On tasks with real learnable signal, models with more clustered representations gain roughly $3.7\times$ more from adaptation than already-spread models. Directly measuring changes in each model's representational geometry during fine-tuning shows that more collapsed frozen representations tend to reorganize more. This reorganization does not consistently lead to higher accuracy, and the relationship varies across the three adaptation mechanisms. Domain-matched pretraining is therefore not sufficient for model selection: general-purpose models with no domain-specific training outperform domain-specialist models on their own home tasks by up to 7%.

1 INTRODUCTION

Representation learning usually means learning a useful representation of data for downstream task. However, it can also be studied independently of task performance, as an analysis of how a system organizes its inputs internally, a structural property that is distinct from whether that organization is useful for a particular task. When independently trained systems converge to similar relational structures despite being constructed from different architectures, objectives, and datasets, it suggests that those structures may be induced by regularities in the underlying phenomenon rather than being arbitrary artifacts of a particular training process (Huh et al., 2024). Conversely, divergence provides evidence that multiple representational solutions are possible and that the geometry of learned representations remains sensitive to the conditions under which learning occurs. This question: which aspects of learned structure are consequences of the training objective, and which are the phenomenon of the world that the system is trying to represent, is central to understanding how models work and what they can transfer to new tasks.

Once understood, these representations can guide model selection and adaptation by revealing how choices, such as the training objective (classification, regression, generation), training architecture(CNN, Transformers, Mixed), training domain(speech, music etc.) and training paradigm (masked modeling, contrastive learning, supervised recognition), shape the learned representation. However, the absolute coordinates of an embedding space are largely arbitrary across independently trained models. Thus we look into the relational structure they impose over inputs, whether two models agree that pairs of inputs are similar or dissimilar, despite differences in architecture, parameters, training data, and learning objective (Kriegeskorte et al., 2008).

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.

Audio provides a useful setting for this question because it contains categories with very different levels of inherent structure within a single modality. Speech and music contain strong, discrete patterns that can be learned from data, whereas mechanical fault sounds and underwater acoustics are more closely tied to continuous physical signals and provide less explicit categorical structure. This range allows us to test whether the same factors govern representational agreement across audio domains, or whether their importance changes with the structure of the underlying sound. We examine this directly in Section 4.1. Audio models are typically evaluated through downstream performance on classification, detection, or recognition tasks, with benchmarks such as SUPERB (Yang et al., 2021), HEAR (Turian et al., 2022), ARCH (La Quatra et al., 2024), and BEANS (Hagiwara et al., 2023) providing standardized evaluation settings across a range of audio tasks. However, these benchmarks measure what a representation enables a model to do, not how the representation itself is organized. Audio models differ in architecture, training objective, training paradigm, and training domain. We examine how these factors shape representational geometry across sound domains with different levels of inherent structure.

A related, practically important question arises when models are adapted to new tasks: can a model's not nfrozen representational geometry predict how it will behave under adaptation? This would enable a pre-adaptation diagnostic for model selection, especially valuable in domains where labeled data is scarce and trial-and-error is costly. We ask whether particular properties of frozen representation geometry, measured before any adaptation, are predictive of how much a model stands to gain when adapted to new tasks.

We introduce ECHO: a systematic study of relational representation geometry across nineteen audio models. Using a fixed multi-domain probe set, we characterize the geometry induced over the same set of acoustic stimuli and measure agreement using representational similarity analysis (Kriegeskorte et al., 2008) and centered kernel alignment (Kornblith et al., 2019). We complement these relational measures with intrinsic dimensionality (Ansuini et al., 2019) and alignment-uniformity decomposition of representation quality (Wang & Isola, 2020). Finally, we investigate whether these geometric properties predict downstream transfer performance.

Our contributions are as follows:

- Cross-model geometric agreement is governed by different factors across domains. Training paradigm explains agreement on structured domains, including music, speech, and urban sound, whereas training-distribution breadth explains agreement on texture-like domains, including bird calls, machine faults, and vessel acoustics. This establishes which factor is relevant for comparing model representations within a given domain rather than assuming a single explanation across domains.
- We identify Embedding uniformity, a geometric measure of how evenly representations are distributed in the embedding space, as a pre-adaptation signal of adaptation headroom across audio domains and tasks. This relationship remains after controlling for model size and task difficulty. We further examine the underlying mechanism by directly measuring representational reorganization during fine-tuning.
- The relationship between uniformity and adaptation gain disappears on tasks for which all tested models perform near chance. This defines a precise and falsifiable scope for the diagnostic rather than treating it as a universal predictor of adaptation performance.

ECHO claims that frozen representation geometry can be measured before fine-tuning to estimate how much a model may benefit from adaptation to a new task. The predictive relationship is bounded by task learnability: it is observed on tasks where frozen models achieve meaningful performance above chance, but disappears on tasks where all tested models remain near chance. The cross-model agreement analysis provides a complementary view by identifying whether training paradigm or training-distribution breadth accounts for representational agreement within a given domain. Together, these results provide a bounded and falsifiable account of what frozen representation geometry can predict about downstream adaptation and where that prediction does not hold.

2 Related Work

Representation Convergence and Geometry: The Platonic Representation Hypothesis (PRH) initiates the idea that independently deep learning models tend to learn increasingly similar internal representations of the world, hypothesizing that different architectures, when optimized on sufficient

data and task variation, are driven toward a shared underlying statistical model of reality (Huh et al., 2024). This hypothesis draws on a broader tradition of comparing neural network representations using second-order similarity structures, such as Centered Kernel Alignment (CKA) (Kornblith et al., 2019). Subsequent literature has mapped the geometric sensitivities of these metrics, illustrating how linear alignment scores can respond to functionality-preserving transformations such as activation scaling or feature space shifts, without altering a model's underlying behavior (Davari et al., 2022; 2023).

Most prior studies of representational convergence have focused on vision and language. To our knowledge, no prior work has systematically measured convergence across a broad, paradigm-diverse set of audio foundation models spanning multiple training paradigms, architectures, and domains. More broadly, convergence analyses have mainly described static representational similarity, without quantifying how representation geometry changes during fine-tuning or using frozen representations to predict downstream adaptation gain before fine-tuning.

Audio Models and Evaluation: Audio models span a wide array of domain specializations; producing highly heterogeneous representations. Here, heterogeneity refers to the differences in how audio is encoded into features like input frontends (1D raw waveforms vs. 2D spectrograms) and distinct architecture types (Transformers, Conformers, CNNs), each imposing distinct inductive biases on the resulting representation. Across this diverse landscape, training paradigms fall into four primary categories, adapted from the self-supervised learning family taxonomy of (Balestriero et al., 2023): mask modeling & reconstruction and joint embedding & self-distillation (the latter merging their separate self-distillation and canonical-correlation families) are drawn from their framework, while supervised and discriminative learning and cross-modal contrastive learning are our own additions and relabeling, extending the scheme to the fully-supervised and cross-modal models in our roster that fall outside a purely self-supervised taxonomy. To evaluate these heterogeneous feature spaces, numerous benchmarks have been proposed: SUPERB (Yang et al., 2021) evaluates speech models across tasks like speaker identification and keyword spotting; HEAR (Turian et al., 2022) covers general environmental and bioacoustic classification; and specialized suites like ARCH (La Quatra et al., 2024) and BEANS (Hagiwara et al., 2023) assess music and animal vocalization tasks.

However, existing benchmarks primarily report task-specific probe accuracy for individual model-task pairs, without examining whether models trained under different objectives develop similar representations. Separately, frozen kNN evaluation is a common training-free proxy for representation quality in individual audio self-supervised works (Fonseca et al., 2021). Neither approach examines whether models trained with different objectives develop similar representations. More broadly, existing evaluations do not connect frozen representation geometry to the amount a model gains from adaptation, leaving adaptation headroom largely unstudied.

Transferability Estimation: These methods aim to predict how well a pretrained model will perform on a new task prior to full fine-tuning. LogME (You et al., 2021) estimates transferability by calculating the maximum marginal likelihood of target labels under a linear probe on frozen features. LEEP (Nguyen et al., 2020) computes expected log-likelihood by measuring how source task predictions align with the target label distribution, while information-theoretic approaches such as NCE (Tran et al., 2019) and H-Score (Bao et al., 2019) evaluate conditional entropy and feature covariance. A parallel line of work evaluates frozen-feature probes as transferability estimators. Renggli et al. (2022) study model selection from growing model repositories using a regret framework and show that combining task-agnostic signals, such as source-task performance, with task-aware signals, such as kNN or linear evaluation on a reference task, outperforms either strategy alone. In a broader comparison of 23 transferability metrics across 16 datasets, Sorkhei et al. (2025) find that simple kNN evaluation predicts transfer accuracy better than purpose-built metrics. These metrics enable computationally efficient backbone selection without costly fine-tuning runs, though systematic benchmark evaluations reveal that many suffer from significant instability and sensitivity to feature scale across diverse architectures (Ibrahim et al., 2022).

However, supervised probability-based estimators like LEEP & NCE require a model to output conditional class probabilities over a known source label set. This becomes a problem because self-supervised architectures produce unconstrained latent embeddings rather than categorical probability distributions, and cannot be evaluated without fitting ad-hoc proxy classifiers or restricting comparison strictly to supervised backbones. The kNN-based estimators above avoid this limitation

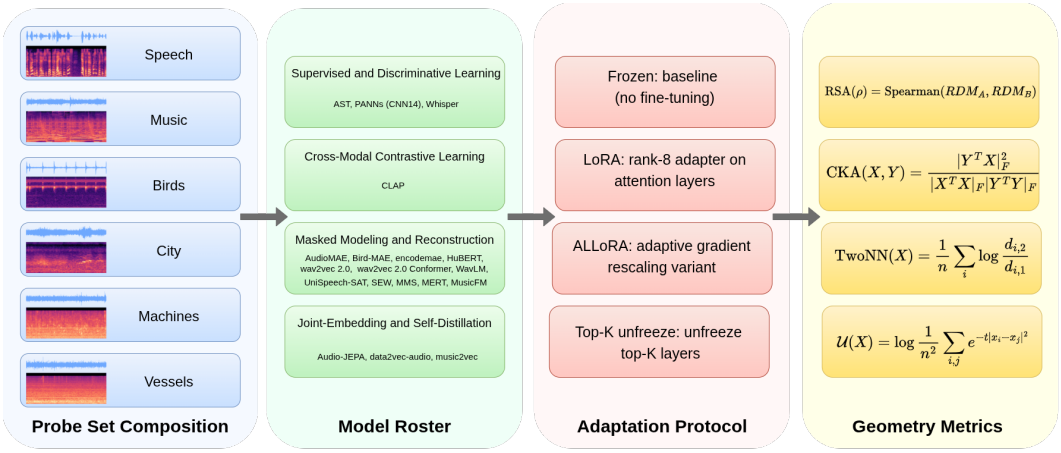


Figure 1: The ECHO Setup. The probe set contains audio clips from six categories: speech, music, birds, city sounds, machine sounds, and vessels. They are processed by 19 independently trained audio models spanning supervised and discriminative learning, cross-modal contrastive learning, masked modeling and reconstruction, and joint-embedding or self-distillation objectives. For each model, we evaluate the frozen representation and its behavior under matched adaptation using LoRA, ALLoRA, or top-K layer unfreezing. We characterize the resulting representations using representational similarity analysis (RSA), centered kernel alignment (CKA), TwoNN intrinsic dimension, and uniformity each computed within a domain's own clip set, never across domains.

by using embedding distances instead of class probabilities. They remain task-aware, however, requiring a labeled reference task, and have so far been evaluated only on vision transfer benchmarks. More broadly, current transferability frameworks treat performance as a static end-state outcome, conflating a model's frozen baseline with its optimization capacity. Consequently, they offer limited insight into how intrinsic, unsupervised feature geometry dictates representational reorganization under downstream adaptation.

3 Experimental Setup

Figure 1 provides an overview of the setup. Let $\mathcal{D} = \{\mathcal{D}_d\}_{d=1}^{6}$ denote the probe set, where each domain-specific subset $\mathcal{D}_d = \{x_i^{(d)}\}_{i=1}^{N_d}$ contains N_d audio clips drawn from domain $d \in \{\text{speech, music, bird, city, machine, vessel}\}$, with N_d varying by domain (Table 1). Each clip $x_i^{(d)}$ is passed through each of the nineteen independently pretrained audio models $m \in \mathcal{M}$ to obtain a latent representation $z_i^{(m,d)} = f_m(x_i^{(d)})$. Rather than directly comparing the raw representations of different models, we compare the relationships between samples in their representation spaces. These relationships are evaluated using two families of representation geometry metrics: pairwise relational metrics $\mathcal{G}^{\text{pair}} = \{g_l : \mathcal{M} \times \mathcal{M} \times \{1, \ldots, 6\} \to \mathbb{R}\}_{l=1}^{2}$ (RSA, CKA) and per-model metrics $\mathcal{G}^{\text{single}} = \{g_l : \mathcal{M} \times \{1, \ldots, 6\} \to \mathbb{R}\}_{l=1}^{2}$ (uniformity, TwoNN), each capturing a different aspect of the global or local structure of the representations. We then examine whether these pre-adaptation geometric properties are related to downstream adaptation. For each eligible model and task t, the model is adapted using Low-Rank Adaptation (LoRA), ALLoRA, or top-K layer unfreezing. This produces an adapted encoder f_m^{adapted} and an adaptation gain $\Delta_{m,t} = A_{m,t}^{\text{adapted}} - A_{m,t}^{\text{frozen}}$, where $A_{m,t}^{\text{frozen}}$ and $A_{m,t}^{\text{adapted}}$ denote performance before and after adaptation, respectively. By keeping the probe set $\mathcal{D}$ fixed across models, ECHO follows three further stages: model evaluation, parameter adaptation, and geometric analysis.

3.1 Probe Set Composition

The probe set $\mathcal{D}_{\text{probe}}$ is partitioned into six acoustic domains $\mathcal{D} = \{d_k\}_{k=1}^{6}$; each drawing from a distinct source dataset and production environment; full statistics are given in Table 1. Each

Domain	Source	Clips	# Classes	SR	Skew (max : min)
Speech	Panayotov et al. (2015)	10,026	251	16 kHz	1.5:1
Music	Defferrard et al. (2017)	8,000	8	44.1 kHz	1:1
City noise	Salamon et al. (2014)	8,732	10	8–192 kHz	2.7:1
Bird sounds	Cañas et al. (2025)	6,335	50	32 kHz	7.2:1
Machine noise	Purohit et al. (2019)	2,400	2	16 kHz	1:1
Vessel	Delarue et al. (2018)†	2,463	6	—	—

† Labels and per-class distribution are subject to a data-sharing agreement and cannot be disclosed.

Table 1: Dataset scale per domain. Skew is the largest-to-smallest class clip ratio. Vessel SR is withheld under the data-sharing agreement.

domain contributes N_k clips, giving $N = \sum_{k=1}^{6} N_k$ total, with N_k ranging from $2{,}400$ (machine noise) to $10{,}026$ (speech). Class balance is not uniform across domains: music and machine noise are balanced by construction (1:1), while bird-sounds (7.2:1) and city noise (2.7:1) imbalances are source-native and preserved. Sample rates span 16 kHz to 8-192 kHz and are not normalised across domains. The vessel domain comprises $2{,}463$ clips from 628 anonymized vessel MMSIs recorded along Canada's East Coast using JASCO AMAR hydrophones Delarue et al. (2018); hydrophone data are archived via the *data.gov* catalogue, but class labels are subject to a data-sharing agreement and cannot be disclosed.

3.2 Model Roster and Adaptation

The nineteen models span four training paradigms. *Supervised and discriminative* models are trained directly on classification or sequence-to-sequence targets: AST (Gong et al., 2021), PANNs (CNN14) (Kong et al., 2020), and Whisper (Radford et al., 2023). *Cross-modal contrastive* learning is represented by CLAP (Elizalde et al., 2023), which aligns audio and text embeddings via a multimodal contrastive loss. *Masked modeling and reconstruction* is the largest group: AudioMAE (Huang et al., 2022), Bird MAE (Rauch et al., 2025), EnCodecMAE (Pepino et al., 2023), HuBERT (Hsu et al., 2021), wav2vec 2.0 (Baevski et al., 2020), wav2vec 2.0 Conformer (Srivastava et al., 2022), WavLM (Chen et al., 2022a), UniSpeech SAT (Chen et al., 2022b), SEW (Wu et al., 2022), MMS (Pratap et al., 2023), MERT (Li et al., 2024), and MusicFM (Won et al., 2024). *Joint-embedding and self-distillation* models predict latent representations without tokenisation or reconstruction: Audio-JEPA (Tuncay et al., 2025), data2vec-audio (Baevski et al., 2022), and music2vec (Li et al., 2022).

For adaptation, fourteen models receive both LoRA (Hu et al., 2022) and ALLoRA (Huang et al., 2024). LoRA is injected into attention query and value projections; the exact target-module names differ by architecture :- `q_proj`/`v_proj` for the eleven standard HuggingFace wav2vec2-style models, `linear_q`/`linear_v` for wav2vec 2.0 Conformer's relative-position attention, fused `qkv` for Bird-MAE's ViT-style blocks, and `query`/`value` scoped to `audio_model.*` for CLAP to avoid injecting into its unused text tower. Four models are ineligible for LoRA and are instead adapted via top-K layer unfreezing: PANNs (CNN14) contains no attention layers and is adapted through its convolutional blocks; Audio-JEPA, AudioMAE, and MusicFM each embed bespoke, non-HuggingFace preprocessing directly into their feature extraction path, which precludes standard module-name injection. EnCodecMAE is excluded from every adaptation condition because its feature extraction is non-differentiable as released; it participates in the frozen-probe geometry evaluation only. Every downstream comparison is performed within the same adaptation condition, comparing LoRA with LoRA, ALLoRA with ALLoRA, and top-K with top-K, without combining results across adaptation methods.

3.3 Representation Geometry Metrics

All geometry metrics use one fixed-size embedding per clip. For most models, this is obtained by mean-pooling the hidden-state sequence over time; AST, CLAP, PANNs, and Bird-MAE use their native clip-level pooling heads. This gives an embedding matrix $Z^{(m)} \in \mathbb{R}^{n \times d_m}$ for each model.

Representational Similarity Analysis (RSA): For each model m, we construct a representational dissimilarity matrix $\mathrm{RDM}^{(m)} \in \mathbb{R}^{n \times n}$. Each entry measures the correlation distance between two

clip embeddings: $1 - r(z_i^{(m)}, z_j^{(m)})$. We then compare two models by computing the Spearman rank correlation between the upper triangles of their RDMs:

$$\rho_{AB} = \mathrm{Spearman}\Big(\mathrm{RDM}_{\triangle}^{(A)},\ \mathrm{RDM}_{\triangle}^{(B)}\Big). \tag{1}$$

Centered Kernel Alignment (CKA): We use it as a secondary measure because prior work has shown it can be sensitive to differences in model architecture, width, and depth (Davari et al., 2023; 2022). We compute linear CKA (Kornblith et al., 2019). Let $H = I_n - \frac{1}{n}ee^\top$, $e = (1, \ldots, 1)^\top$, be the centering matrix. The centered Gram matrices are $\tilde{G}_A = H(Z^{(A)}Z^{(A)\top})H$ and $\tilde{G}_B = H(Z^{(B)}Z^{(B)\top})H$. CKA is then computed as

$$\mathrm{CKA}(A, B) = \frac{\|\tilde{G}_A^\top \tilde{G}_B\|_F^2}{\|\tilde{G}_A\|_F \, \|\tilde{G}_B\|_F}. \tag{2}$$

Uniformity: Following Wang & Isola (Wang & Isola, 2020), uniformity measures how evenly the embeddings are distributed on the unit hypersphere. We first L2-normalise each embedding to obtain $\hat{z}_i$, then compute

$$\mathcal{U}(Z) = \log \frac{1}{n(n-1)} \sum_{i \neq j} e^{-2\,\|\hat{z}_i - \hat{z}_j\|^2}, \tag{3}$$

The sum includes all pairs of different clips. Lower, more negative values indicate that the embeddings are more spread out, while higher values indicate that they are concentrated in a smaller region.

Intrinsic Dimensionality (TwoNN): We estimate the intrinsic dimensionality of each model's representation space using the TwoNN estimator (Facco et al., 2017), implemented with `skdim` (Bac et al., 2021). For each clip i, let $d_{i,1}$ and $d_{i,2}$ be the Euclidean distances to its first and second nearest neighbours in the raw embedding space. The estimate is

$$\widehat{\mathrm{ID}}(Z) = \left(\frac{1}{n} \sum_{i=1}^{n} \log \frac{d_{i,2}}{d_{i,1}} \right)^{-1}. \tag{4}$$

Higher values indicate that the representations occupy more effective dimensions, while lower values indicate that they lie on a more compressed, lower-dimensional structure. TwoNN is known to be sensitive to the density and class composition of the point set it is computed on (Schulte & Rügamer, 2026; Osman et al., 2026); we confirmed this on our own data. We therefore use TwoNN only to compare models within a single domain's fixed probe set, never to compare raw magnitudes across domains.

4 Results

4.1 Finding 1: Cross-model agreement shifts along an acoustic spectrum

We asked whether agreement between independently trained audio encoders is better organized by training domain & paradigm (d & p) or by the breadth (whether their pretraining data were domain-specific or mixed) of the training distribution. For each acoustic domain, we compared the RSA matrix with a same-group matrix $S_g(i, j) = \mathbf{1}\{g(i) = g(j)\}$, where g denotes one of the two taxonomies. The resulting Mantel correlation, $r_g = \mathrm{Pearson}(\mathrm{RSA}_{\triangle}, S_{g,\triangle})$, is positive when models assigned to the same group have more similar representational geometry than models assigned to different groups. Significance was evaluated by permuting model identities 20,000 times. Across the twelve domain–taxonomy tests, we controlled the false discovery rate with Benjamini–Hochberg correction at $q = 0.05$, retaining sorted p-values that satisfied $p_{(k)} \leq (k/12)q$.

Across domains, the relative alignment varies along a continuous spectrum rather than being discrete (Fig. 2). This raises the possibility that the relative importance of training paradigm versus training breadth depends systematically on the target sound, though we do not have an independent variable to test this ordering directly, and the present result should be read as descriptive rather than as evidence of a graded relationship. The downstream analyses test whether these differences in frozen representations are reflected in linear-probe accuracy, measuring information learned by a classifier, and k-nearest-neighbour accuracy, measuring how closely samples from the same class are grouped in the embedding space.

4.2 Finding 2: Discriminative training and domain relevance shape frozen usability

We showcase what makes a frozen embedding usable by evaluating each encoder two ways: a linear probe and k-nearest-neighbour. On music, city and speech a discriminatively trained model ranks at or above second place on both metrics. On city, AST reaches 0.835 linear / 0.844 kNN, while AudioMAE, trained by reconstruction, reaches 0.733 linear but only 0.612 kNN. This pattern was challenged by a direct falsification test, an earlier analysis based on PANNs and Bird-MAE suggested that supervised paradigms trade raw separability for head-exploitability. AST, however, shows no such trade-off and achieves the highest BirdCLEF kNN accuracy. The narrower claim is that discriminative training pressure, whether contrastive or supervised, produces embeddings that are both raw-separable and head-exploitable. This holds across all tasks tested, kNN-to-linear accuracy ratio is higher for discriminative than reconstruction models (sign test, $p = 0.0625$).

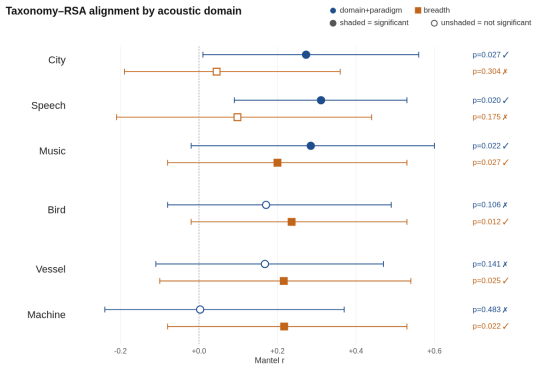


Figure 2: Taxonomy–RSA alignment across acoustic domains. Points show the Mantel correlation between pairwise RSA and whether two models belong to the same domain & paradigm group (blue) or breadth group (orange). Horizontal bars show bootstrap 95% confidence intervals across models. Filled markers indicate permutation significance at $p < 0.05$; open markers indicate non-significance.

Moreover,when we keep paradigm and architecture fixed like AudioMAE and Bird-MAE only differ in training domain, Bird-MAE beats AudioMAE by $\times 1.61$ linear and $\times 1.91$ kNN accuracy on BirdCLEF, an advantage that vanishes on music and machine(within ± 0.03), confirming it is domain-specific. Both effects converge on a single geometric quantity that how evenly embeddings are spread across the representation space (Wang & Isola, 2020) predicts kNN accuracy directly (pooled $\rho = -0.358$, $p = 0.0004$) and predicts the kNN–linear gap even more tightly ($\rho = -0.380$, $p = 0.0002$). More uniform embeddings yield higher kNN accuracy without any downstream training. Contrastive training directly optimizes this quantity as part of its loss (Wang & Isola, 2020); AST reaches similarly high uniformity through supervised classification instead, showing that different optimization paths can converge on the same geometric outcome. Appendix A.4 reports this relationship is not attributable to model size.

4.3 Finding 3: Frozen geometry predicts adaptation headroom through measurable representational reorganization

Finding 4.2 showed that embedding uniformity governs whether a model's frozen distances are already usable, without any downstream training. We test whether the same quantity also predicts how much a model improves when it is fine-tuned. Fine-tuning each encoder and measuring gain as the change in accuracy from a frozen linear probe to the adapted encoder, three contributors together explain most of the variation: task difficulty, embedding uniformity, and domain match. Splitting music/city/speech into top and bottom thirds by uniformity within each fine-tuning seed, the most-clustered third gains $3.72 \pm 0.16\times$ more on average than the most-spread third, the source of the abstract's ratio. Fitting one linear model across every (model, task, seed) row gives gain $= 0.098 + 0.50 \cdot$ difficulty $+ 0.10 \cdot$ uniformity $+ 0.13 \cdot$ domain-match $+$ residual(all coefficients $p < 0.001$). Difficulty is the normalized distance of a domain's mean frozen accuracy from chance toward perfect accuracy, averaged across all models. This relationship remains robust after controlling for model size (Appendix A.4) and without requiring any labels, achieves predictive performance comparable to LogME and superior to k-nearest-neighbour evaluated in a label-efficient regime (Appendix A.5). Aggregating this decomposition by the previously introduced taxonomy group and training paradigm, and further splitting the residual into model-identity, task-identity,

seed, and cell-offset components via a two-way fixed-effects fit, gives the full picture of what drives adaptation gain across the roster:

Group	Training Paradigm	n	Difficulty (%)	Uniformity (%)	Domain-match (%)	Model-ID (%)	Task-ID (%)	Seed (%)	Cell-offset (%)
D&P	Supervised/discriminative	3	19.3	11.6	3.8	28.5	11.8	10.1	14.8
D&P	Cross-modal contrastive	1	27.2	65.8	0.5	0.0	5.9	0.3	0.3
D&P	Masked modeling/reconstruction	11	21.0	6.4	7.4	42.1	0.9	21.7	0.5
D&P	Joint-embedding/self-distillation	3	19.8	1.8	3.9	17.1	0.9	39.7	16.7
Both†	Supervised/discriminative	3	8.4	70.8	1.1	12.1	0.0	2.6	5.1
Both†	Cross-modal contrastive	1	4.2	92.8	0.5	0.0	0.0	0.9	1.5
Both†	Masked modeling/reconstruction	11	8.4	11.3	11.7	13.8	0.0	3.3	51.6
Both†	Joint-embedding/self-distillation	3	7.7	0.4	18.9	11.7	0.0	1.6	59.8
Breadth	Supervised/discriminative	3	26.3	16.5	0.8	4.8	22.8	13.0	15.9
Breadth	Cross-modal contrastive	1	28.6	13.3	0.8	0.0	2.3	19.6	35.4
Breadth	Masked modeling/reconstruction	11	30.7	9.6	2.5	13.6	13.2	30.3	0.1
Breadth	Joint-embedding/self-distillation	3	46.8	6.9	1.4	6.1	4.7	25.2	8.9

Table 2: Relative squared attribution share of each gain contributor, by taxonomy group and training paradigm. n=1/n=3 rows are single- or near-single-model estimates, read descriptively; only the n=11 masked-modeling/reconstruction row per group is large enough to trust quantitatively. †Both is single-task (music only), so task-identity is structurally 0%. Seed reflects seed-to-seed variability plus unmodeled model$\times$task interaction.

The Breadth decomposition reveals a contrast between difficulty and uniformity. Difficulty is the largest contributor in every Breadth cell but one, reaching 46.8% for joint-embedding/self-distillation models. This is consistent with Finding 4.1, where these three domains were governed primarily by how much learnable signal a task provides rather than by the taxonomy of the model. Uniformity shows the opposite pattern: its contribution is sharply concentrated, reaching 92.8% for CLAP alone in the Both group and 70.8% for the small discriminative cluster in the same group, while remaining in the low across every Breadth cell. This concentration reassures that uniformity's role is specific to cross-model structure rather than a general driver of individual models' adaptation gains.

Uniformity contributes across the decomposition, but it is dominant only in the discriminative/contrastive cells noted above. TwoNN intrinsic dimension (Facco et al., 2017), computed on the same embeddings, shows no relationship with adaptation gain on five of the six domains tested ($|\rho| \leq 0.25$, $p \geq 0.4$). The exception is the vessel domain, where TwoNN correlates positively with gain under both LoRA ($\rho = +0.644$, permutation $p = 0.016$) and ALLoRA ($\rho = +0.486$, $p = 0.085$), with the direction reproduced across four of five fine-tuning seeds ($\rho = 0.31$–0.60) and one seed near zero. Unlike uniformity, whose predictive power is concentrated on tasks with real learnable signal and is null on vessel, TwoNN shows the reverse pattern: its relationship with adaptation gain appears specifically on the near-chance domain where uniformity does not. We tested two candidate explanations for this reversal: raw embedding collapse, measured by uniformity, and the sensitivity of TwoNN to sparse per-category sampling. Neither accounts for the observed pattern. Vessel uniformity is not associated with TwoNN values ($\rho = 0.135$, $p = 0.587$), and domains with substantially sparser per-class sampling than vessel do not show a comparable effect. The mechanism therefore remains unresolved, but these analyses rule out two plausible explanations and narrow the set of factors that could account for the reversal.

We measure representational geometry change during adaptation using self-RSA, the Spearman correlation between the frozen and adapted RDM upper triangles computed on the same clips. Lower self-RSA indicates greater reorganization; higher self-RSA indicates that frozen geometry is largely preserved. Because uniformity directly predicts self-RSA, a bivariate association between self-RSA and gain could reflect uniformity's joint influence on both variables rather than a genuine reorganization mechanism. Table 3 therefore reports both the bivariate and partial correlations between self-RSA and gain for each domain-by-adaptation-method cell, where the partial correlation controls for uniformity.

The mechanism is not universal, and its two links differ in stability. The first link is broadly consistent: across most domain-by-method cells, models with more collapsed frozen representations undergo greater representational reorganization during adaptation (Table 3, Column 4). The second link, from reorganization to accuracy gain, is more selective. After controlling for uniformity, the partial correlation between self-RSA and gain is correctly signed and significant in only four of eighteen cells. Thus, frozen geometry is a consistent predictor of how much reorganization occurs

		Self-RSA mechanism		
Domain	Condition	Uniformity → self-RSA	Bivariate ρ	Partial ρ (ctrl. unif.)
Speech	LoRA	$\rho = -0.441,\ p = 0.0016$	-0.429	-0.172 ($p = 0.242$)
Speech	ALLoRA	$\rho = -0.410,\ p = 0.0006$	-0.343	-0.124 ($p = 0.348$)
Speech	Top-K	$\rho = -0.742,\ p = 0.0005$	-0.943	$\mathbf{-0.874}$ ($p < 0.001$)
Music	LoRA	$\rho = -0.275,\ p = 0.024$	-0.444	$\mathbf{-0.363}$ ($p = 0.003$)
Music	ALLoRA	$\rho = -0.252,\ p = 0.039$	-0.481	$\mathbf{-0.428}$ ($p < 0.001$)
Music	Top-K	$\rho = -0.349,\ p = 0.138$	-0.295	$+0.117$ ($p = 0.631$)
City	LoRA	$\rho = -0.121,\ p = 0.341$	-0.367	$\mathbf{-0.321}$ ($p = 0.008$)
City	ALLoRA	$\rho = -0.202,\ p = 0.110$	-0.253	-0.209 ($p = 0.090$)
City	Top-K	$\rho = -0.776,\ p < 0.001$	-0.365	$+0.592$ ($p = 0.009$)
Bird	LoRA	$\rho = -0.378,\ p = 0.001$	$+0.122$	$+0.164$ ($p = 0.180$)
Bird	ALLoRA	$\rho = -0.491,\ p < 0.001$	$+0.106$	$+0.213$ ($p = 0.087$)
Bird	Top-K	$\rho = -0.969,\ p < 0.001$	-0.747	$+0.029$ ($p = 0.904$)
Machine	LoRA	$\rho = -0.281,\ p = 0.023$	-0.061	-0.108 ($p = 0.390$)
Machine	ALLoRA	$\rho = -0.300,\ p = 0.015$	-0.100	-0.130 ($p = 0.306$)
Machine	Top-K	$\rho = +0.078,\ p = 0.745$	-0.265	-0.252 ($p = 0.297$)
Vessel	LoRA	$\rho = -0.302,\ p = 0.011$	-0.117	-0.141 ($p = 0.249$)
Vessel	ALLoRA	$\rho = -0.237,\ p = 0.048$	$+0.184$	$+0.160$ ($p = 0.183$)
Vessel	Top-K	$\rho = -0.682,\ p = 0.001$	-0.084	$+0.037$ ($p = 0.878$)

Table 3: Self-RSA mechanism check, per-seed granularity.

during adaptation, but whether that reorganization improves accuracy depends on the domain and adaptation method.

On speech, the partial association remains significant only for top-K unfreezing ($\rho = -0.874$, $p < 0.001$); under LoRA and ALLoRA, the bivariate association disappears after controlling for uniformity. On music, the partial correlation remains significant under LoRA and ALLoRA ($\rho = -0.363$ and -0.428, respectively), but not under top-K. On city, it remains significant under LoRA ($\rho = -0.321$, $p = 0.008$), while top-K shows a sign reversal after controlling for uniformity. On bird and machine, the uniformity→self-RSA link remains negative and significant across adaptation methods, whereas the partial self-RSA→gain correlation is consistently null. Thus, collapsed representations can undergo substantial reorganization during adaptation without producing an accuracy gain when the downstream task provides limited learnable signal.

Together, these results separate two effects that a single uniformity–gain correlation cannot distinguish. The amount of geometric reorganization is largely determined by the model's frozen starting geometry, whereas whether that reorganization produces an accuracy gain depends on the domain and adaptation method. A model can therefore undergo substantial geometric change during adaptation without improving accuracy.

5 Conclusion

Section 4.3 shows what determines adaptation gain in practice. Task difficulty accounts for roughly four times as much of adaptation gain as uniformity, making it the most informative quantity to check before fine-tuning. When the probe performs above chance, uniformity provides additional information for distinguishing otherwise plausible models. Uniformity is computed directly from unlabeled embeddings and requires no additional labels or reference task. More collapsed representations are associated with greater representational reorganization during adaptation, which we measure directly under three different adaptation mechanisms. When the frozen probe performs at chance, this relationship disappears, and the other geometric measures we tested also provide little information about adaptation gain.

These results define the scope of the proposed diagnostic. It predicts adaptation headroom, or how much a model may gain from fine-tuning, rather than its final adapted performance for audio. The taxonomy underlying Finding 4.1 was also developed incrementally rather than pre-registered. TwoNN shows a domain-specific association in the vessel data that remains unexplained. These limits are part of the result: task difficulty is the strongest predictor of adaptation gain, geometry provides additional information when sufficient learnable signal is present, and the geometric signal disappears when that signal is limited.

AI USE STATEMENT

All research ideas and hypotheses in this work are original to the authors. Generative AI tools were used for initial code production, and for running and monitoring parallel experiments. We also used generative AI tools to produce initial analysis reports, which we then used as a starting point for further drafting and revision. We have reviewed all AI-assisted work. The final version of this paper is written, understood, narrated, and explained by us in our own understanding. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses a confidential AIS-derived vessel acoustic dataset in addition to standard public audio benchmarks. Vessel identifiers (MMSI) in the confidential dataset are anonymized prior to any analysis, and no vessel-identifying information is reported, released, or reproducible from any result in this paper. This dataset was used under a research agreement restricting its use to non-commercial research; we do not release it. All other datasets used are publicly available under their respective licenses. This work does not involve human subjects research.

REPRODUCIBILITY STATEMENT

Our probe set, model roster, adaptation protocol, and geometry metric formulas are fully specified in Section 3 (Experimental Setup), including exact clip counts and sources per domain (Table 1), the full model list with training paradigm and adaptation eligibility (Table 4, Appendix A.2), and closed-form definitions of RSA, CKA, uniformity, and TwoNN (Eqs. 1–4). Hardware, software versions, hyperparameters for LoRA/ALLoRA/top-K adaptation, and per-domain evaluation splits are given in Appendix A.1. Every accuracy number underlying Finding 4.3 and the attribution decomposition is reported in full in Appendix A.3 (mean$\pm$std across seeds and folds, per model, domain, and condition), and the full cross-model RSA structure, pretrained and finetuned, is reported in Appendix A.6.

One artifact is not distributable, the self-RSA and the frozen-vs-adapted representational reorganization analysis (Section 4.3) require a full pairwise representational dissimilarity matrix (RDM) per model, domain, adaptation condition, and seed, computed over each domain's complete clip pool (up to $\sim$38,000 clips pooled across the six domains). Each RDM is a dense $N \times N$ matrix in floating point, and the complete set of per-clip adapted-RDM checkpoints used in this work totals approximately 512 GB on disk, well beyond what can be hosted as supplementary material. The vessel domain's checkpoints are additionally restricted from public release under the governing data-sharing agreement (Ethics Statement), independent of size. We therefore release the code to regenerate these RDMs from the public datasets and pretrained checkpoints listed in Table 4, together with every aggregate result (accuracy, RSA/CKA matrices, self-RSA values) that was computed from them, rather than the intermediate RDM files themselves. Code is available at `https://anonymous.4open.science/r/ECHO_AUDIO/README.md`.

REFERENCES

Alessio Ansuini, Alessandro Laio, Jakob H. Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Jonathan Bac, Evgeny M. Mirkes, Alexander N. Gorban, Ivan Tyukin, and Andrei Zinovyev. Scikit-dimension: A Python package for intrinsic dimension estimation. *Entropy*, 23(10):1368, 2021. doi: 10.3390/e23101368.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 12449–12460, 2020.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning (ICML)*, pp. 1298–1312, 2022.

Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, Avi Schwarzschild, Andrew Gordon Wilson, Jonas Geiping, Quentin Garrido, Pierre Fernandez, Amir Bar, Hamed Pirsiavash, Yann LeCun, and Micah Goldblum. A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210*, 2023. URL https://arxiv.org/abs/2304.12210.

Yajie Bao, Yang Li, Shao-Lun Huang, Lin Zhang, Lizhong Zheng, Amir Roshan Zamir, and Leonidas J. Guibas. An information-theoretic approach to transferability in task transfer learning. In *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 2309–2313, Taipei, Taiwan, September 2019. IEEE. doi: 10.1109/ICIP.2019.8803726. URL https://doi.org/10.1109/ICIP.2019.8803726.

Juan Sebastian Cañas, Stefan Kahl, Tom Denton, et al. Overview of BirdCLEF+ 2025: Multi-taxonomic sound identification in the middle magdalena, Colombia. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025)*, volume 4038 of *CEUR Workshop Proceedings*, pp. 2909–2919, 2025. URL https://ceur-ws.org/Vol-4038/.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and Furu Wei. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022a.

Sanyuan Chen, Yu Wu, Chengyi Wang, Zhengyang Chen, Zhuo Chen, Shujie Liu, Jian Wu, Yao Qian, Furu Wei, Jinyu Li, and Xiangzhan Yu. UniSpeech-SAT: Universal speech representation learning with speaker-aware pre-training. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6152–6156. IEEE, 2022b. doi: 10.1109/ICASSP43922.2022.9747077. URL https://doi.org/10.1109/ICASSP43922.2022.9747077.

MohammadReza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. On the inadequacy of CKA as a measure of similarity in deep learning. In *ICLR Workshop on Geometrical and Topological Representation Learning*, 2022. URL https://openreview.net/forum?id=rK841rby6xc.

MohammadReza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. Reliability of CKA as a similarity measure in deep learning. In *The Eleventh International Conference on Learning Representations (ICLR)*. OpenReview.net, 2023. URL https://openreview.net/forum?id=8HRvyxc606.

Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. FMA: A dataset for music analysis. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2017. arXiv:1612.01840.

Julien Delarue, Kimberley Kowarski, Emily Maxner, Jessica MacDonnell, and Bruce Martin. Acoustic monitoring along Canada's east coast: August 2015 to July 2017. Technical Report 01279, JASCO Applied Sciences for the Environmental Studies Research Fund, Dartmouth, NS, Canada, 2018. Version 1.0. 120 pp + appendices. Available via JASCO Applied Sciences and the *data.gov* catalogue.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP: Learning audio concepts from natural language supervision. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, Rhodes Island, Greece, June 2023. IEEE. doi: 10.1109/ICASSP49357.2023.10095889. URL https://doi.org/10.1109/ICASSP49357.2023.10095889. arXiv preprint arXiv:2206.04769.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighbourhood information. *Scientific Reports*, 7:12140, 2017. doi: 10.1038/s41598-017-11873-y.

Eduardo Fonseca, Diego Ortego, Kevin McGuinness, Noel E. O'Connor, and Xavier Serra. Unsupervised contrastive learning of sound event representations. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 371–375. IEEE, 2021. doi: 10.1109/ICASSP39728.2021.9415009.

Yuan Gong, Yu-An Chung, and James Glass. AST: Audio spectrogram transformer. In *Interspeech 2021*, pp. 571–575. ISCA, 2021. doi: 10.21437/Interspeech.2021-698. URL https://doi.org/10.21437/Interspeech.2021-698.

Masato Hagiwara, Benjamin Hoffman, Jen-Yu Liu, Maddie Cusimano, Felix Effenberger, and Katie Zacarian. BEANS: The benchmark of animal sounds. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5. IEEE, 2023. doi: 10.1109/ICASSP49357.2023.10096686. URL https://doi.org/10.1109/ICASSP49357.2023.10096686.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

Hai Huang, Randall Balestriero, et al. ALLoRA: Adaptive learning rate mitigates LoRA fatal flaws. *arXiv preprint arXiv:2410.09692*, 2024.

Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 28708–28720, 2022.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642. PMLR, 2024. URL https://proceedings.mlr.press/v235/huh24a.html. arXiv preprint arXiv:2405.07987.

Shibal Ibrahim, Natalia Ponomareva, and Rahul Mazumder. Newer is not always better: Rethinking transferability metrics, their peculiarities, stability and performance. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, volume 13713 of *Lecture Notes in Computer Science*, pp. 693–709. Springer, 2022.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020. doi: 10.1109/TASLP.2020.3030497. URL https://doi.org/10.1109/TASLP.2020.3030497.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019. URL http://proceedings.mlr.press/v97/kornblith19a.html. arXiv preprint arXiv:1905.00414.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008.

Moreno La Quatra, Alkis Koudounas, Lorenzo Vaiani, Elena Baralis, Luca Cagliero, Paolo Garza, and Sabato Marco Siniscalchi. Benchmarking representations for speech, music, and acoustic events. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pp. 505–509. IEEE, 2024. doi: 10.1109/ICASSPW62465.2024.10625960. URL https://doi.org/10.1109/ICASSPW62465.2024.10625960. arXiv preprint arXiv:2405.00934.

Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Chenghua Lin, Xingran Chen, Anton Ragni, Hanzhi Yin, Zhijie Hu, Haoyu He, Emmanouil Benetos, Norbert Gyenge, Ruibo Liu, and Jie Fu. MAP-Music2Vec: A simple and effective baseline for self-supervised music audio representation learning. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR)*, 2022.

Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Chenghua Lin, Anton Ragni, Emmanouil Benetos, Norbert Gyenge, Roger B. Dannenberg, Ruibo Liu, et al. MERT: Acoustic music understanding model with large-scale self-supervised training. In *International Conference on Learning Representations (ICLR)*, 2024.

Cuong V. Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. LEEP: A new measure to evaluate transferability of learned representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7294–7305. PMLR, 2020. URL https://proceedings.mlr.press/v119/nguyen20b.html.

Arwa Osman, Marco Baroni, and Iuri Macocco. Local and global regimes of geometric complexity in language model representations. *arXiv preprint arXiv:2608.14361*, 2026. doi: 10.48550/arXiv.2608.14361. URL https://arxiv.org/abs/2608.14361.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. LibriSpeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210. IEEE, 2015. doi: 10.1109/ICASSP.2015.7178964. URL https://doi.org/10.1109/ICASSP.2015.7178964.

Leonardo Pepino, Pablo Ernesto Riera, and Luciana Ferrer. EnCodecMAE: Leveraging neural codecs for universal audio representation learning. *arXiv preprint arXiv:2309.07391*, 2023. URL https://arxiv.org/abs/2309.07391.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Aravind Vyas, Maryam Fazel-Zarandi, et al. Scaling speech technology to 1,000+ languages. *arXiv preprint arXiv:2305.13516*, 2023.

Harsh Purohit, Ryo Tanabe, Kenji Ichige, Takashi Endo, Yuki Nikaido, Kaori Suefusa, and Yohei Kawaguchi. MIMII dataset: Sound dataset for malfunctioning industrial machine investigation and inspection. In *Proceedings of the 4th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, pp. 209–213, 2019. arXiv:1909.09347.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning (ICML)*, pp. 28492–28518, 2023.

Lukas Rauch, René Heinrich, Ilyass Moummad, Alexis Joly, Bernhard Sick, and Christoph Scholz. Can masked autoencoders also listen to birds? *arXiv preprint arXiv:2504.12880*, 2025.

Cedric Renggli, André Susano Pinto, Luka Rimanic, Joan Puigcerver, Carlos Riquelme, Ce Zhang, and Mario Lučić. Which model to transfer? finding the needle in the growing haystack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9205–9214, 2022.

Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM International Conference on Multimedia*, pp. 1041–1044. ACM, 2014. doi: 10.1145/2647868.2655045. URL https://doi.org/10.1145/2647868.2655045.

Rickmer Schulte and David Rügamer. Rethinking intrinsic dimension estimation in neural representations. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, 2026. doi: 10.48550/arXiv.2604.20276. URL https://arxiv.org/abs/2604.20276. Accepted for publication.

Moein Sorkhei, Christos Matsoukas, Johan Fredin Haslum, Emir Konuk, and Kevin Smith. k-nn as a simple and effective estimator of transferability. *Transactions on Machine Learning Research*, 2025.

Sangeeta Srivastava, Yun Wang, Andros Tjandra, Anurag Kumar, Chunxi Liu, Kritika Singh, and Yatharth Saraf. Conformer-based self-supervised learning for non-speech audio tasks. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8862–8866. IEEE, 2022. doi: 10.1109/ICASSP43922.2022.9747613. URL https://doi.org/10.1109/ICASSP43922.2022.9747613.

Anh T. Tran, Cuong V. Nguyen, and Tal Hassner. Transferability and hardness of supervised classification tasks. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1395–1405. IEEE, 2019. doi: 10.1109/ICCV.2019.00148. URL https://doi.org/10.1109/ICCV.2019.00148.

Ludovic Tuncay, Etienne Labbé, Emmanouil Benetos, and Thomas Pellegrini. Audio-JEPA: Joint-embedding predictive architecture for audio representation learning. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. to appear. IEEE, 2025. URL https://hal.science/hal-05128180v1/file/Audio-JEPA.pdf. Preprint also available as arXiv:2507.02915.

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*, PMLR, 2022.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

Minz Won, Yun-Ning Hung, and Duc Le. A foundation model for music informatics. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1226–1230. IEEE, 2024. doi: 10.1109/ICASSP48485.2024.10448314. URL https://doi.org/10.1109/ICASSP48485.2024.10448314.

Felix Wu, Kwangyoun Kim, Jing Pan, Kyu Jeong Han, Kilian Q. Weinberger, and Yoav Artzi. SEW: Squeezed and efficient wav2vec with disentangled attention. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7667–7671. IEEE, 2022. doi: 10.1109/ICASSP43922.2022.9747432. URL https://doi.org/10.1109/ICASSP43922.2022.9747432.

Shu-Wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-Tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-Yi Lee. SUPERB: Speech processing universal PERformance benchmark. In *Interspeech 2021*, pp. 1194–1198. ISCA, 2021. doi: 10.21437/Interspeech.2021-1775. URL https://doi.org/10.21437/Interspeech.2021-1775. arXiv preprint arXiv:2105.01051.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. LogME: Practical assessment of pre-trained models for transfer learning. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12133–12143. PMLR, 2021. URL https://proceedings.mlr.press/v139/you21b.html.

A Appendix

A.1 Implementation and Training Details

All experiments run on four GPU configurations available across two clusters: full NVIDIA H100 cards, MIG-partitioned H100 fractions (1g.10 GB, 2g.20 GB, and 3g.40 GB slices), and two NVIDIA RTX PRO 6000 Blackwell Max-Q GPUs. The primary software environment uses Python 3.11.5, PyTorch 2.11.0, CUDA 12.9, and cuDNN 9.1. BirdNET-family inference (where

applicable to model loading) runs in an isolated CPU-only environment (Python 3.11.4, TensorFlow 2.15.1) to avoid dependency conflicts with the primary PyTorch stack.

LoRA and ALLoRA are trained for up to 8 epochs with learning rate 10^{-4}, batch size 16, AdamW, and early stopping at patience 3 on held-out validation accuracy. LoRA uses rank $r = 8$, $\alpha = 16$, and dropout 0.05 uniformly across all eligible models. ALLoRA uses rank 8 and $\eta = 8.0$; we note that $\eta^2 = 64$ lies outside the range evaluated in the original paper ($\eta^2 \in \{1, 2, 4\}$) — the value matches LoRA's α/r ratio and was not separately ablated. Top-K unfreezing ($K = 2$ blocks) uses the same training budget. The frozen linear-probe baseline uses Adam, learning rate 10^{-3}, 30 epochs, and patience 5. All stochastic processes (model initialization, data-order shuffling, and train/validation splitting) are repeated across 5 seeds.

Evaluation splits are domain-specific: FMA (music) uses an 80/10/10 random split; LibriSpeech (speech) uses a per-speaker-stratified 70/15/15 split; UrbanSound8K (city) uses a 4-fold subset of its native 10-fold scheme; BirdCLEF (bird) uses 5 species-balanced folds; MIMII (machine) uses 4 leave-one-machine-type-out folds; and the vessel domain uses 5 seeded folds under the governing data-sharing agreement. All conditions are evaluated on plain classification accuracy. Fold-based domains (city, bird, machine) select one fixed representative fold to define the adapted encoder used for geometry computation (self-RSA, RDM, CKA), rather than averaging RDMs across folds, since a geometry computation requires one fixed embedding space per model.

A.2 Model Roster and Adaptation Eligibility

Table 4 lists all nineteen models with their training paradigm and adaptation eligibility, as described in Section 3.2.

Model	Training Paradigm	Adaptation
AST	Supervised/discriminative	LoRA, ALLoRA
PANNs (CNN14)	Supervised/discriminative	Top-K only
Whisper	Supervised/discriminative	LoRA, ALLoRA
CLAP	Cross-modal contrastive	LoRA, ALLoRA
AudioMAE	Masked modeling/reconstruction	Top-K only
Bird-MAE	Masked modeling/reconstruction	LoRA, ALLoRA
EnCodecMAE	Masked modeling/reconstruction	Frozen only
HuBERT	Masked modeling/reconstruction	LoRA, ALLoRA
wav2vec 2.0	Masked modeling/reconstruction	LoRA, ALLoRA
wav2vec 2.0 Conformer	Masked modeling/reconstruction	LoRA, ALLoRA
WavLM	Masked modeling/reconstruction	LoRA, ALLoRA
UniSpeech-SAT	Masked modeling/reconstruction	LoRA, ALLoRA
SEW	Masked modeling/reconstruction	LoRA, ALLoRA
MMS	Masked modeling/reconstruction	LoRA, ALLoRA
MERT	Masked modeling/reconstruction	LoRA, ALLoRA
MusicFM	Masked modeling/reconstruction	Top-K only
Audio-JEPA	Joint-embedding/self-distillation	Top-K only
data2vec-audio	Joint-embedding/self-distillation	LoRA, ALLoRA
music2vec	Joint-embedding/self-distillation	LoRA, ALLoRA

Table 4: Full model roster (19 models), grouped by training paradigm (Section 3.2). Fourteen models receive both LoRA and ALLoRA; four are adapted only via top-K layer unfreezing (no attention layers, or bespoke non-differentiable preprocessing that precludes standard LoRA module injection); EnCodecMAE participates in frozen-probe geometry only, since its feature extraction is non-differentiable as released. LoRA/ALLoRA target-module naming differs by architecture family (Section 3.2): `q_proj`/`v_proj` for the eleven standard HuggingFace wav2vec2-style models (AST, Whisper, HuBERT, wav2vec 2.0, WavLM, UniSpeech-SAT, SEW, MMS, MERT, data2vec-audio, music2vec), `linear_q`/`linear_v` for wav2vec 2.0 Conformer, fused `qkv` for Bird-MAE, and `query`/`value` scoped to `audio_model.*` for CLAP.

Correlation	n	Raw ρ	Partial ρ (controlling $\log n_{\text{params}}$)	p	VIF
Uniformity vs. kNN accuracy (Finding 4.2)	94	-0.358	-0.341	$7.7\text{e}{-4}$	1.02
Uniformity vs. adaptation gain (Finding 4.3)	108	$+0.447$	$+0.443$	$1.6\text{e}{-6}$	1.02

Table 5: Spearman partial correlations controlling for $\log(n_{\text{params}})$, using the same populations as the corresponding headline results.

Metric	n	Pooled ρ vs. gain	p
kNN (100 labeled clips/class)	90	-0.059	0.581
LogME (You et al., 2021)	90	$+0.429$	$2.4\text{e}{-5}$
Uniformity (Ours)	90	$+0.448$	$9.5\text{e}{-6}$

Table 6: Uniformity against two label-efficient transferability estimators as predictors of adaptation gain, on the same population.

A.3 Full Accuracy Table

Table 7 reports frozen and adapted accuracy for every model, domain, and adaptation condition used in this study, as mean$\pm$std across the 5 training seeds (Appendix A.1); for city, bird, and machine, the fold-level results are first averaged within each seed before averaging across seeds.

A.4 Parameter-Count Robustness Check

The geometric correlations in Sections 4.2 and 4.3 could be confounded by model size. We test this by computing Spearman partial correlations controlling for $\log(n_{\text{params}})$ and the variance inflation factor (VIF) for uniformity and $\log(n_{\text{params}})$. Each analysis uses the same population as its corresponding headline result: the Finding 4.2 check pools 19 models across 5 domains (fma-genre, urbansound8k, birdclef, librispeech, mimii; $n = 94$, with one model–domain pair missing its kNN row), while the Finding 4.3 check pools 18 models across all six domains, including vessel ($n = 108$ model$\times$domain pairs, seed-averaged). Partial correlations are computed via rank-residualization: both variables are rank-transformed, each is linearly residualized against the rank of $\log(n_{\text{params}})$, and the Spearman correlation is recomputed on the residuals. Parameter counts use each model's encoder only, as reported in Table 4. The resulting partial correlations and VIF values are reported in Table 5.

Both correlations change only slightly after controlling for parameter count, with VIF ≈ 1.02 in both cases. The uniformity relationships in Findings 4.2 and 4.3 therefore remain after accounting for model size.

A.5 Comparison Against Label-Efficient Transferability Estimators

Uniformity requires no labels, motivating a comparison with transferability estimators that use limited label supervision. We compare uniformity against two established estimators on the same tasks: k-nearest-neighbour accuracy using up to 100 labeled reference clips per class ($k = 5$, cosine similarity, evaluated on the held-out remainder of each domain's clip pool), and LogME (You et al., 2021), a closed-form estimator of transfer performance from frozen features and labels that requires no fine-tuning. All three measures are computed from the frozen embeddings for the five public domains with per-clip labels available (fma-genre, urbansound8k, birdclef, librispeech, mimii; vessel is excluded because per-clip labeled embeddings for the confidential domain were never extracted outside the official fine-tuning pipeline). Each estimator is correlated with the same adaptation gain used throughout Section 4.3 ($n = 90$, 18 models $\times$ 5 domains). Results are reported in Table 6.

Uniformity matches or exceeds both baselines while requiring no labels. The kNN baseline shows no association with adaptation gain and has the opposite sign. The difference is significant (Steiger's $Z = 3.14$, $p = 0.0017$, for two overlapping correlations sharing the gain variable). This result is compatible with (Sorkhei et al., 2025), who found that kNN can predict transfer accuracy effectively. Their analysis concerns the adapted performance level, while our analysis measures adaptation gain, defined as the change from frozen to adapted performance. These quantities capture different prop-

erties: frozen kNN accuracy can estimate eventual adapted accuracy without indicating how much performance will be gained through adaptation. Our results support this distinction and align with the broader point in Section 2 that frozen performance and optimization capacity represent distinct aspects of transferability. Against LogME, uniformity achieves $\rho = 0.448$ compared with $\rho = 0.429$ for LogME, with no statistically significant difference (Steiger's $Z = 0.15$, $p = 0.88$). The two estimators therefore provide comparable predictions of adaptation gain, with uniformity requiring no labels, reference task, or classifier fitting. Across domains, uniformity has the strongest correlation among the three predictors on fma-genre ($\rho = 0.734$), urbansound8k ($\rho = 0.674$), and librispeech ($\rho = 0.783$). On mimii, all three predictors are weak or null, with uniformity at $\rho = -0.057$, consistent with the near-chance behavior described in Finding 4.3. Birdclef provides the main exception among domains with a measurable signal, where uniformity has a comparatively weak correlation ($\rho = 0.216$), consistent with the exception noted in Section 4.3.

A.6 Cross-Model RSA Structure: Pretrained vs. Finetuned

Figure 3(a) presents the complete 19×19 RSA matrix for the pretrained (frozen) models, averaged across all six domains. Each panel uses its own observed minimum and maximum off-diagonal RSA values for the colour scale, rather than the nominal $[-1, 1]$ correlation range, since the observed RSA values occupy a much narrower range. Diagonal self-correlations ($= 1$) therefore saturate the upper end of each panel's scale. Panels (b) and (c) show the corresponding finetuned matrices for the 14 models eligible for LoRA and ALLoRA, averaged across all six domains and five seeds. Panel (d) shows the corresponding matrix for the four models eligible for top-K. As described in Section 3.2, the LoRA/ALLoRA and top-K conditions apply to disjoint model subsets, while EnCodecMAE remains frozen. A single finetuned matrix covering all 19 models would therefore mix adaptation conditions, so the eligible subsets are shown separately.

The strongest agreement in the pretrained matrix is concentrated within recognizable training-paradigm groups. HuBERT–WavLM ($\rho = 0.694$) and HuBERT–UniSpeech-SAT ($\rho = 0.661$) are masked-modeling speech models trained on the same 960-hour LibriSpeech corpus. MERT–MusicFM ($\rho = 0.659$) forms the corresponding pair for masked-modeling music models, while AST–PANNs ($\rho = 0.697$) is the strongest pair among supervised/discriminative general-audio models. AST–CLAP ($\rho = 0.638$) is the strongest cross-paradigm pair, consistent with Finding 4.2, which identifies discriminative training pressure as a shared organizing factor across contrastive and supervised objectives. The weakest agreement pairs are concentrated around data2vec-audio and AudioMAE. data2vec-audio–music2vec ($\rho = 0.149$), despite belonging nominally to the same joint-embedding/self-distillation family, ranks among the five lowest-agreement pairs in the roster. AudioMAE forms the other extreme, with AudioMAE–data2vec-audio ($\rho = 0.133$) and AudioMAE–wav2vec2 ($\rho = 0.142$). data2vec-audio's isolation persists across domains and also appears in the finetuned matrices, where it forms four of the five lowest-agreement pairs under both LoRA and ALLoRA.

For the common 14-model subset, mean off-diagonal RSA increases from 0.365 in the pretrained matrix to 0.416 under LoRA and 0.418 under ALLoRA. Thus, in this domain-pooled view, models show somewhat greater representational agreement after adaptation. This comparison is descriptive and averages across all six domains and both adaptation conditions. It does not establish whether the increase is consistent across domains or driven by a subset of them, which is outside the scope of this appendix.

Model	Speech				Music				City				Bird				Machine				Vessel			
	Fz	LoRA	ALoRA	TopK	Fz	LoRA	ALoRA	TopK	Fz	LoRA	ALoRA	TopK	Fz	LoRA	ALoRA	TopK	Fz	LoRA	ALoRA	TopK	Fz	LoRA	ALoRA	TopK
AST	0.81±0.00	0.99±0.00	0.99±0.00	–	0.55±0.00	0.57±0.01	0.56±0.01	–	0.86±0.01	0.89±0.02	0.89±0.02	–	0.36±0.05	0.50±0.05	0.49±0.05	–	0.52±0.07	0.50±0.03	0.52±0.04	–	0.21±0.04	0.30±0.07	0.32±0.08	–
PANNs (CNN14)	0.39±0.00	–	–	0.84±0.11	0.54±0.01	–	–	0.56±0.01	0.80±0.03	–	–	0.84±0.02	0.17±0.06	–	–	0.45±0.05	0.49±0.08	–	–	0.53±0.07	0.16±0.05	–	–	0.18±0.06
Whisper	0.04±0.04	0.82±0.06	0.92±0.02	–	0.35±0.04	0.52±0.02	0.50±0.01	–	0.27±0.09	0.77±0.03	0.77±0.03	–	0.05±0.02	0.30±0.06	0.31±0.05	–	0.52±0.05	0.51±0.05	0.53±0.05	–	0.16±0.02	0.30±0.10	0.30±0.09	–
CLAP	0.81±0.01	0.87±0.01	0.90±0.01	–	0.57±0.01	0.57±0.01	0.58±0.01	–	0.77±0.02	0.88±0.03	0.88±0.03	–	0.11±0.05	0.21±0.05	0.23±0.05	–	0.46±0.06	0.51±0.04	0.50±0.03	–	0.13±0.05	0.29±0.11	0.27±0.09	–
AudioMAE	0.15±0.01	–	–	0.97±0.01	0.41±0.01	–	–	0.51±0.01	0.37±0.10	–	–	0.83±0.03	0.04±0.02	–	–	0.47±0.05	0.48±0.04	–	–	0.49±0.03	0.18±0.09	–	–	0.34±0.15
Bird-MAE	0.62±0.01	0.96±0.01	0.96±0.01	–	0.44±0.01	0.49±0.02	0.50±0.01	–	0.66±0.04	0.86±0.02	0.85±0.03	–	0.11±0.07	0.43±0.05	0.39±0.04	–	0.51±0.03	0.50±0.06	0.48±0.04	–	0.22±0.05	0.32±0.06	0.32±0.08	–
EnCodecMAE	0.98±0.00	–	–	–	0.43±0.03	–	–	–	0.64±0.04	–	–	–	0.09±0.05	–	–	–	0.50±0.05	–	–	–	0.22±0.06	–	–	–
HuBERT	0.74±0.00	0.92±0.01	0.92±0.01	–	0.34±0.03	0.36±0.01	0.38±0.03	–	0.52±0.05	0.63±0.04	0.61±0.04	–	0.08±0.04	0.23±0.04	0.21±0.04	–	0.47±0.07	0.46±0.07	0.47±0.07	–	0.14±0.04	0.20±0.09	0.26±0.08	–
wav2vec 2.0	0.01±0.00	0.28±0.12	0.21±0.19	–	0.13±0.01	0.32±0.02	0.35±0.04	–	0.13±0.03	0.38±0.04	0.41±0.06	–	0.03±0.01	0.08±0.02	0.09±0.03	–	0.50±0.02	0.42±0.06	0.44±0.07	–	0.20±0.07	0.25±0.07	0.22±0.06	–
wav2vec 2.0 Conf.	0.41±0.00	0.74±0.05	0.72±0.03	–	0.29±0.02	0.34±0.01	0.35±0.01	–	0.51±0.03	0.52±0.04	0.53±0.03	–	0.08±0.04	0.18±0.04	0.17±0.04	–	0.45±0.05	0.48±0.01	0.51±0.02	–	0.18±0.08	0.31±0.14	0.27±0.12	–
WavLM	0.43±0.01	0.64±0.03	0.71±0.02	–	0.32±0.00	0.42±0.01	0.42±0.02	–	0.49±0.05	0.51±0.05	0.50±0.04	–	0.07±0.04	0.16±0.05	0.15±0.04	–	0.48±0.06	0.54±0.09	0.54±0.07	–	0.23±0.13	0.18±0.04	0.26±0.11	–
UniSpeech-SAT	0.75±0.01	0.98±0.00	0.98±0.00	–	0.35±0.00	0.44±0.01	0.43±0.02	–	0.48±0.05	0.66±0.05	0.66±0.03	–	0.10±0.05	0.21±0.05	0.21±0.05	–	0.53±0.05	0.47±0.07	0.46±0.06	–	0.15±0.06	0.22±0.08	0.33±0.09	–
SEW	0.26±0.19	0.88±0.02	0.89±0.01	–	0.26±0.04	0.37±0.01	0.36±0.03	–	0.29±0.06	0.61±0.03	0.61±0.03	–	0.04±0.02	0.16±0.04	0.17±0.05	–	0.49±0.05	0.48±0.07	0.48±0.07	–	0.22±0.05	0.21±0.10	0.23±0.06	–
MMS	0.34±0.14	0.93±0.01	0.93±0.01	–	0.27±0.07	0.42±0.02	0.43±0.01	–	0.35±0.08	0.61±0.02	0.65±0.05	–	0.05±0.02	0.19±0.04	0.20±0.05	–	0.47±0.05	0.46±0.05	0.48±0.04	–	0.16±0.07	0.18±0.09	0.32±0.05	–
MERT	0.65±0.00	0.99±0.00	0.99±0.00	–	0.41±0.01	0.53±0.01	0.53±0.01	–	0.61±0.04	0.78±0.03	0.79±0.03	–	0.08±0.04	0.32±0.06	0.32±0.05	–	0.47±0.03	0.50±0.05	0.51±0.05	–	0.18±0.10	0.33±0.13	0.35±0.14	–
MusicFM	0.80±0.00	–	–	0.68±0.01	0.49±0.00	–	–	0.54±0.00	0.66±0.02	–	–	0.73±0.03	0.15±0.06	–	–	0.16±0.05	0.51±0.06	–	–	0.47±0.04	0.16±0.06	–	–	0.32±0.14
Audio-JEPA	0.20±0.00	–	–	0.76±0.01	0.37±0.01	–	–	0.44±0.01	0.34±0.03	–	–	0.67±0.01	0.04±0.02	–	–	0.10±0.04	0.50±0.03	–	–	0.52±0.04	0.17±0.08	–	–	0.31±0.10
data2vec-audio	0.10±0.01	0.84±0.01	0.83±0.01	–	0.28±0.03	0.38±0.01	0.37±0.02	–	0.20±0.05	0.49±0.03	0.53±0.04	–	0.03±0.01	0.06±0.02	0.08±0.02	–	0.54±0.06	0.48±0.05	0.49±0.06	–	0.20±0.05	0.17±0.08	0.18±0.08	–
music2vec	0.14±0.00	0.15±0.01	0.17±0.01	–	0.44±0.01	0.55±0.01	0.56±0.01	–	0.32±0.04	0.68±0.04	0.69±0.03	–	0.03±0.02	0.11±0.03	0.10±0.03	–	0.51±0.05	0.47±0.04	0.50±0.03	–	0.22±0.06	0.28±0.10	0.28±0.10	–

Table 7: Accuracy (mean±std across seeds, and across folds for city/bird/machine) for every model, domain, and adaptation condition. Fz = frozen linear probe; LoRA/ALoRA/TopK = adapted accuracy under that condition. $--$ indicates the model is not eligible for that adaptation condition (Section 3.2): 14 models receive LoRA/ALoRA, 4 receive top-K only, and EnCodecMAE is frozen-only.

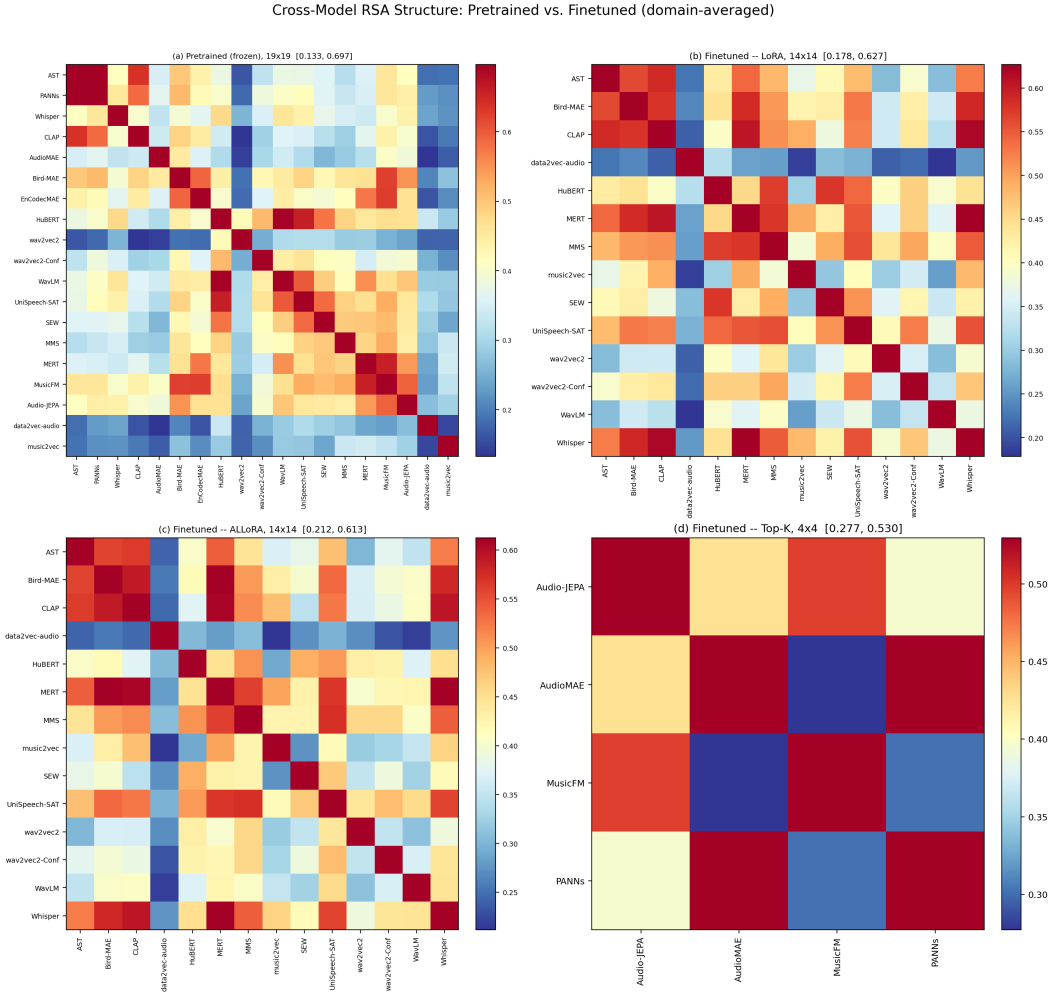


Figure 3: Cross-model RSA structure, domain-averaged. (a) Pretrained (frozen), full 19×19 roster. (b) Finetuned under LoRA, restricted to the 14 LoRA-eligible models, averaged across seeds. (c) Finetuned under ALLoRA, same 14 models. (d) Finetuned under top-K unfreezing, restricted to the 4 top-K-eligible models. Each panel's colour range is set to its own observed off-diagonal min/max (shown in the panel title), not the nominal $[-1, 1]$ correlation range.