THE NO FREE LUNCH OF SELF-SUPERVISED BRAIN MODELS: BEHAVIOR PREDICTION, LEARNING OBJECTIVE, AND DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep neural networks provide leading quantitative models of sensory systems. By circumventing the need for massive labeled datasets, self-supervised pretrained models are increasingly viewed as a free lunch for sensory system modeling. In this study, we evaluate this assumption by applying four categories of computational models—four acoustic, three semantic, three supervised, and eight self-supervised models—to predict human auditory cortical responses and perceptual behavior. Our results demonstrate that this lunch is far from free. In auditory cortex, model-class advantages shifted across the cortical hierarchy: supervised models generally showed stronger alignment in primary and secondary regions, whereas selected self-supervised models surpassed the strongest supervised baselines by approximately 25–30% in tertiary regions. Within the self-supervised class, cortical predictivity varied substantially across learning objectives and pretraining datasets, including among closely matched model families. This cortical advantage did not extend to perceptual behavior, where the strongest supervised DNN exceeded the strongest self-supervised model by approximately 54% and 57% in the sound- and word-judgment tasks, respectively. Self-supervised models also tended to peak at earlier network depths. Together, these findings reveal distinct trade-offs between cortical and behavioral alignment and highlight the importance of learning objective and pretraining data in sensory system modeling.

1 INTRODUCTION

The rapid proliferation of high-performing deep neural networks (DNNs) has established them as promising computational models of sensory systems. These models offer a framework for studying how different inductive biases shape the emergent alignment between artificial and biological representations (Yamins & DiCarlo, 2016). Self-supervised learning has become increasingly attractive relative to supervised learning, both for its engineering performance and because it avoids reliance on large-scale human-provided labels. Supervised training typically relies on massive labeled datasets and is therefore unlikely to provide a complete account of biological learning, as human and animal perception develops largely in real-world environments where explicit labels are limited. But is self-supervised learning all we need to model sensory systems? To address this question in the auditory domain, we examine how model–human alignment varies across behavioral and cortical targets, learning objectives, and pretraining data.

Sensory systems form hierarchical processing stages. Early-stage neurons represent relatively simple features, such as tone frequencies, whereas higher-order regions represent increasingly complex properties such as speech. Similar hierarchical organizations emerge naturally in DNNs (Yamins et al., 2014). Because neural data in many benchmark studies are collected from subjects during passive viewing or listening (Schrimpf et al., 2018), strong neural alignment does not necessarily require representations optimized for the behavioral task under study. Predicting human behavior, however, introduces a distinct challenge: behavior requires sensory representations to support a particular decision or judgment. While a self-supervised model can capture rich statistical structure in sensory inputs, it is not explicitly optimized for the task-specific representational boundaries required by a particular behavioral readout. This raises the possibility that self-supervised repre-

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.

sentations may align well with cortical responses without showing a corresponding advantage in task-dependent perceptual behavior.

The alignment between a model's internal representations and biological responses may also depend on its learning objective. Self-supervised representation learning encompasses substantially different objectives. In audio, Wav2Vec2 (Baevski et al., 2020) uses contrastive learning, HuBERT (Hsu et al., 2021) employs masked prediction of discrete hidden units, and WavLM (Chen et al., 2022) further incorporates denoising into masked prediction. These objectives impose different constraints on the information preserved in learned representations, motivating us to examine how learning objective is associated with auditory cortical alignment.

Pretraining data provide a second potential source of variation. Widely used self-supervised speech models are often pretrained on relatively curated corpora such as LibriSpeech (Panayotov et al., 2015), whereas other available corpora differ substantially in language, duration, acoustic content, and recording conditions. Previous work suggests that environmental statistics can shape learned sensory representations (Mehrer et al., 2021; Francl & McDermott, 2022). We therefore compare self-supervised models pretrained on several distinct auditory corpora to examine how cortical predictivity varies across pretraining datasets. Because these corpora differ along multiple dimensions simultaneously, these comparisons characterize dataset-associated differences rather than isolating the causal contribution of any single property of the pretraining data.

Our main contributions are fourfold: (1) showing that model-class advantages shift across the auditory cortical hierarchy, with selected self-supervised models showing their strongest advantage over supervised baselines in tertiary auditory cortex; (2) demonstrating that this cortical advantage does not extend to human perceptual behavior, where supervised models generally show stronger predictive alignment and later layer-wise peaks; (3) revealing that the relationship between learning objective and cortical predictivity varies across pretraining corpora; and (4) showing that changing the pretraining corpus can substantially shift cortical alignment within closely matched self-supervised model families across the auditory hierarchy.

2 Related Work

Supervised models of auditory perception. Task-optimized DNNs have been widely used to model human auditory behavior. Supervised auditory networks have reproduced aspects of sound recognition (Kell et al., 2018), as well as spatial hearing (Francl & McDermott, 2022; Saddler & McDermott, 2024), pitch perception (Saddler et al., 2021), and selective auditory attention (Griffith et al., 2026). These studies further suggest that both task optimization and the statistics of the training environment can influence the correspondence between model and human behavior (Francl & McDermott, 2022). However, the behavioral consequences of self-supervised auditory representations remain less well characterized.

Supervised models of auditory cortex. Supervised DNNs have also provided successful computational models of auditory cortex, with representations evaluated against both fMRI (Güçlü et al., 2016; Kell et al., 2018; Millet & King, 2021; Giordano et al., 2023; Tuckute et al., 2023) and intracranial recordings (Keshishian et al., 2020; Mischler et al., 2023; Ahmed et al., 2025; Norman-Haignere et al., 2025; Rupp et al., 2025). These studies have shown that task-optimized auditory networks can account for substantial neural variance and can exhibit correspondence between model depth and stages of the auditory cortical hierarchy. In particular, Kell et al. (2018) demonstrated that supervised task optimization can produce representations predictive of both auditory behavior and cortical responses.

Self-supervised models of auditory cortex. More recently, self-supervised speech models such as Wav2Vec2 and HuBERT have emerged as strong candidate models of auditory cortical representations (Millet et al., 2022; Vaidya et al., 2022; Li et al., 2023). These studies demonstrate that substantial neural alignment can emerge without explicit auditory category supervision. However, prior work has primarily focused on cortical responses, leaving two questions less clear: whether strong cortical alignment in self-supervised models extends to human perceptual behavior, and how cortical predictivity varies across learning objectives and pretraining corpora. Here, we address these questions by comparing behavioral and cortical alignment across supervised and self-supervised models

and by examining closely matched self-supervised model families that differ in learning objective or pretraining corpus (Fig. 1).

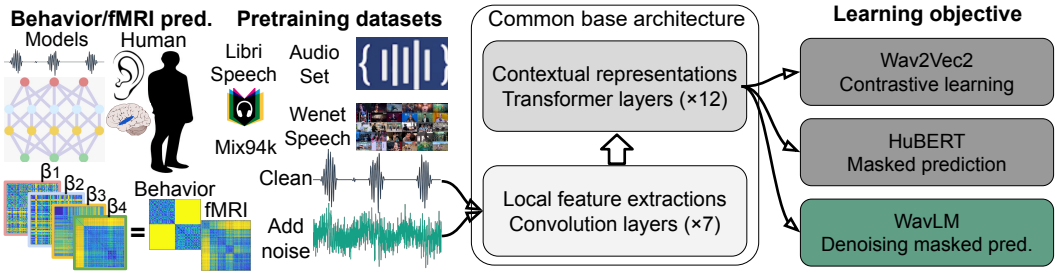

Figure 1: Overview of the study framework. For each human dataset, the corresponding stimulus set was presented to participants and processed by the computational models. Model-component representational dissimilarity matrices (RDMs) were combined using cross-validated linear regression to predict behavioral and fMRI RDMs. Self-supervised models were compared across different pretraining corpora and three learning objectives: contrastive learning (Wav2Vec2), masked prediction (HuBERT), and masked prediction with denoising (WavLM).

3 METHODS

3.1 A BENCHMARK OF AUDITORY MODELS

We analyzed 18 audio models spanning four acoustic, three semantic, three supervised, and eight self-supervised models (Table 1; full specifications are provided in Appendix A). The acoustic models capture spectrotemporal and psychoacoustic properties of sound, whereas the semantic models represent descriptive verb-and-noun labels using pretrained text embeddings. The three supervised CNN-based models (Kell, VGGish, and YamNet) were pretrained for word, music-genre, or sound-event classification.

The eight self-supervised models comprise Wav2Vec2, HuBERT, and WavLM variants pretrained on four corpora: AudioSet, LibriSpeech, Mix94k, and WenetSpeech (Appendix A.4). These models support two complementary comparisons. First, models pretrained on the same corpus were compared across closely matched model families with different learning objectives: contrastive learning in Wav2Vec2, masked prediction in HuBERT, and masked prediction with denoising in WavLM. Second, models within the same family and learning objective but pretrained on different corpora were compared to characterize how cortical alignment varies across pretraining datasets. This comparison included LibriSpeech, WenetSpeech, and AudioSet variants of HuBERT and Wav2Vec2, as well as the LibriSpeech and Mix94k variants of WavLM. Because these corpora differ simultaneously in language, duration, acoustic content, recording conditions, and other training characteristics, these comparisons characterize dataset-associated differences rather than isolating the causal contribution of any single property of the pretraining data.

3.2 BEHAVIORAL AND FMRI DATASETS

We analyzed previously collected behavioral and high-field fMRI datasets of natural-sound perception. The behavioral dataset (Giordano et al., 2010) comprised 40 participants divided into two groups of 20. Participants performed hierarchical similarity judgments on 80 natural sounds (40 living and 40 nonliving), using either the sounds themselves or corresponding verb-and-noun descriptions. Behavioral dissimilarities were derived separately for the sound- and word-judgment tasks. Full experimental details are provided in Appendix B.1.

The fMRI dataset (Santoro et al., 2017) comprised five participants who listened to 288 natural sounds spanning six categories: speech, voice, animal, music, nature, and tools, during 7T fMRI acquisition. Cortical responses were analyzed within six bilateral auditory regions of interest (ROIs): Heschl's gyrus (HG), planum temporale (PT), planum polare (PP), middle superior temporal gyrus (mSTG), posterior superior temporal gyrus (pSTG), and anterior superior temporal gyrus (aSTG).

Table 1: Overview of the acoustic, semantic, supervised, and self-supervised models. Pretraining duration is reported in thousands of hours (k) where applicable. Abbreviations: CoG, cochleagram; NA, not applicable. Wikipedia+ denotes composite text corpora including Wikipedia and other large text or web sources. TIMIT+ denotes the combined training resources used by the Kell model, including TIMIT, Wall Street Journal, and the Million Song Dataset.

Model	Description / Objective	Representation	Pretraining Data
CochleaG	Peripheral spectral analysis	Time-varying / averaged CoG	NA
MTF	Spectrotemporal analysis	Time-varying / averaged MTF	NA
Texture	Sound texture analysis	5 statistical components	NA
AudDims	Psychoacoustic features	5 auditory dimensions	NA
GNewsW2V	Word2Vec embedding	Latent (300)	Google News
GloVe	Word embedding	Latent (300)	Wikipedia+
GUSE	Sentence encoder	Latent (512)	Wikipedia+
Kell	Supervised classification	Word + genre (587 + 41)	TIMIT+ 3k
VGGish	Supervised classification	Audio-event labels	YouTube / AudioSet
YamNet	Supervised classification	AudioSet labels (521)	AudioSet 5.8k
WavLM-LS	Masked pred. + denoising	Latent (768)	LibriSpeech 0.96k
HuBERT-LS	Masked prediction	Latent (768)	LibriSpeech 0.96k
Wav2Vec2-LS	Contrastive learning	Latent (768)	LibriSpeech 0.96k
WavLM-Mix94k	Masked pred. + denoising	Latent (768)	Mix94k 94k
HuBERT-Wenet	Masked prediction	Latent (768)	WenetSpeech 10k
Wav2Vec2-Wenet	Contrastive learning	Latent (768)	WenetSpeech 10k
HuBERT-AS	Masked prediction	Latent (768)	AudioSet 5.8k
Wav2Vec2-AS	Contrastive learning	Latent (768)	AudioSet 5.8k

These ROIs were grouped as primary (HG), secondary (PT and PP), and tertiary (mSTG, pSTG, and aSTG) auditory cortex. Full acquisition and preprocessing details are provided in Appendix B.1.

3.3 Predictive representational similarity analysis framework

To evaluate how well computational models account for human behavioral and neural representations, we used the predictive representational similarity analysis (RSA) framework described by Giordano et al. (2023). Unlike voxel-wise encoding models, which map stimulus features directly to voxel responses (Kell et al., 2018), or conventional RSA, which typically compares representational dissimilarity matrices (RDMs) using rank correlations (Tuckute et al., 2023), predictive RSA learns a linear combination of model RDM components on training data and evaluates its predictions on held-out human data.

Behavioral RDMs were derived from the hierarchical similarity judgments described above. For the fMRI data, ROI-specific RDMs were constructed for each participant using Euclidean distances between demeaned stimulus-wise GLM response patterns. For each computational model component, model RDMs were constructed using pairwise cosine distances between stimulus representations. Further details of RDM construction and predictive analysis are provided in Appendix B.2.

For each RDM, the lower-triangular elements were extracted and flattened into a vector of unique pairwise distances. Let $\mathbf{d}_{\text{train}}^{\text{human}}$ denote the empirical behavioral or fMRI distance vector in the training fold, and $\mathbf{d}_{l,\text{train}}^{\text{model}}$ the corresponding distance vector for model component l. An ordinary least-squares (OLS) regression estimated coefficients β_l that combined the n model components:

$$\mathbf{d}_{\text{train}}^{\text{human}} = \sum_{l=1}^{n} \beta_l \mathbf{d}_{l,\text{train}}^{\text{model}}. \tag{1}$$

The fitted coefficients were then frozen and applied to the corresponding model distances in the test fold:

$$\hat{\mathbf{d}}_{\text{test}}^{\text{human}} = \sum_{l=1}^{n} \beta_l \mathbf{d}_{l,\text{test}}^{\text{model}}. \tag{2}$$

Predictive performance was quantified using the cross-validated coefficient of determination,

$$R^2_{\text{cv}} = 1 - \frac{SSE_{\text{test}}}{SST_{\text{test}}}, \tag{3}$$

where SSE_{test} is the squared prediction error on held-out data and SST_{test} is the total variance of the held-out human distances. Behavioral analyses used 100 random participant split-halves (10 training and 10 test participants per task), whereas fMRI analyses used all 10 exhaustive participant splits (three training and two test participants) together with cross-validation across stimulus sets. Full cross-validation and pairwise dimensionality details are provided in Appendix B.2 and Appendix B.3.

Statistical significance was assessed using 10,000 matrix permutations. Inference was performed on the median R^2_{cv} across cross-validation splits, with a maximum-statistic procedure controlling the family-wise error rate at FWER ≤ 0.05 (Appendix B.2).

3.4 Local topological similarity: mutual k-nearest neighbors

To complement predictive RSA, we quantified local representational similarity using mutual k-nearest-neighbor (MkNN) overlap. For each stimulus, MkNN measures the fraction of its k nearest neighbors that are shared between the biological and model representational spaces, averaged across stimuli. We evaluated MkNN across neighborhood sizes k; the expected overlap for independently organized representations is $k/(N-1)$. Higher values therefore indicate greater preservation of local representational structure. Additional methodological details are provided in Appendix B.4.

4 Results

4.1 Model-class advantages shift across the auditory cortical hierarchy

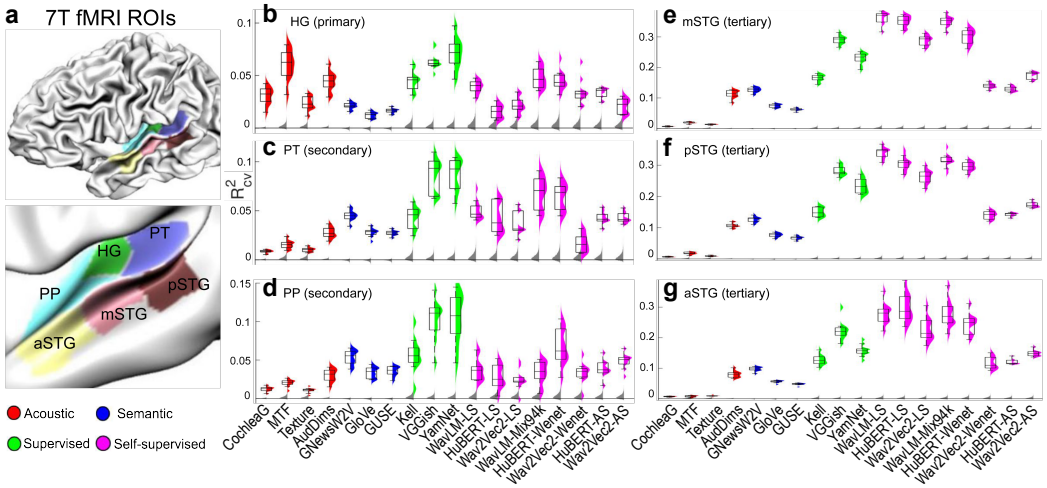

Figure 2: Auditory cortical anatomy and model predictivity across the cortical hierarchy. **a**, Anatomical locations of the six auditory fMRI regions of interest (ROIs): Heschl's gyrus (HG), planum temporale (PT), planum polare (PP), middle superior temporal gyrus (mSTG), posterior superior temporal gyrus (pSTG), and anterior superior temporal gyrus (aSTG). **b–g**, Cross-validated predictive performance (R^2_{cv}) of the acoustic, semantic, supervised, and self-supervised models in HG (**b**), PT (**c**), PP (**d**), mSTG (**e**), pSTG (**f**), and aSTG (**g**). Colors denote model class: acoustic (red), semantic (blue), supervised (green), and self-supervised (magenta). Box plots show the median, interquartile range, and whiskers extending to 1.5 times the interquartile range. Gray elements indicate permutation-based null results. $N = 5$ fMRI participants.

We first examined how well acoustic, semantic, supervised, and self-supervised models predicted representational geometry across the auditory cortical hierarchy (Fig. 2). The six auditory ROIs

comprised primary auditory cortex (HG), secondary auditory cortex (PT and PP), and tertiary auditory cortex (mSTG, pSTG, and aSTG) (Fig. 2a).

Model rankings changed substantially across cortical stages. In primary and secondary auditory cortex (Fig. 2b–d), supervised models generally showed the strongest predictive performance. In HG, YamNet achieved the highest median R^2_{cv} among the evaluated models, while the acoustic MTF model showed higher median predictivity than all eight self-supervised models. In PT and PP, VGGish and YamNet showed the highest median predictive performance among the evaluated models.

A different pattern emerged in tertiary auditory cortex (Fig. 2e–g). Four self-supervised models—WavLM-LS, HuBERT-LS, WavLM-Mix94k, and HuBERT-Wenet—showed higher median R^2_{cv} than all three supervised models in mSTG, pSTG, and aSTG. The highest self-supervised median R^2_{cv} values were 0.346 in mSTG, 0.371 in pSTG, and 0.287 in aSTG, compared with corresponding best supervised values of 0.276, 0.293, and 0.221, respectively. Relative to the strongest supervised baseline, these values correspond to improvements of **25.3%**, **26.6%**, and **29.9%** in mSTG, pSTG, and aSTG, respectively. Semantic models showed relatively low cortical predictivity across all six ROIs.

Together, these results show that model-class advantages vary across the auditory cortical hierarchy. Supervised models show the strongest alignment in primary and secondary auditory cortex, with acoustic models remaining competitive in HG, whereas selected self-supervised models show their clearest advantage over supervised models in tertiary auditory cortex.

4.2 Cortical alignment varies with learning objective and pretraining dataset

Given the substantial variability among self-supervised models in cortical prediction, we next asked whether this variation was associated with learning objective and pretraining dataset (Fig. 3).

We first compared models pretrained on the same corpus but using different self-supervised learning objectives (Fig. 3a). Among the three LibriSpeech-pretrained models, cortical predictivity followed the ordering WavLM-LS $>$ HuBERT-LS $>$ Wav2Vec2-LS in four of the six ROIs, and WavLM-LS showed higher median R^2_{cv} than Wav2Vec2-LS in all six regions. A similar objective-dependent difference was observed for the WenetSpeech-pretrained models, where HuBERT-Wenet showed higher median predictivity than Wav2Vec2-Wenet across all six ROIs. However, this ordering was not universal. Among the AudioSet-pretrained models, Wav2Vec2-AS showed higher median R^2_{cv} than HuBERT-AS in four of the six ROIs, including PP and all three tertiary regions. Thus, the relationship between learning objective and cortical predictivity varies across pretraining corpora rather than following a single objective ranking.

We next compared models within the same model family and learning objective but pretrained on different corpora (Fig. 3b). Within the HuBERT family, HuBERT-Wenet and HuBERT-AS showed higher median predictivity than HuBERT-LS in primary and secondary auditory cortex, whereas HuBERT-LS showed the highest median predictivity in all three tertiary ROIs. Within the Wav2Vec2 family, Wav2Vec2-Wenet showed the highest median predictivity in HG and Wav2Vec2-AS in PT and PP, whereas Wav2Vec2-LS showed the highest median predictivity in all three tertiary ROIs. A related shift was also observed within the WavLM family: WavLM-Mix94k showed higher median predictivity than WavLM-LS in HG and PT, whereas WavLM-LS showed higher predictivity in PP and all three tertiary ROIs.

To visualize the stage-specific representational geometries underlying these differences, we compared PP and aSTG with the corresponding best-predicting HuBERT variants (Fig. 3c,d). HuBERT-Wenet qualitatively resembled the categorical organization observed in PP, whereas HuBERT-LS more closely resembled the distinct organization observed in aSTG. These examples illustrate that pretrained variants within the same model family can preferentially align with different stages of the auditory cortical hierarchy.

Together, these comparisons show that cortical alignment varies with both learning objective and pretraining dataset. Importantly, changing the pretraining corpus does not produce a uniform shift in cortical predictivity; relative model advantages vary across the auditory cortical hierarchy. Because the pretraining corpora also differ in language, duration, acoustic content, recording conditions, and

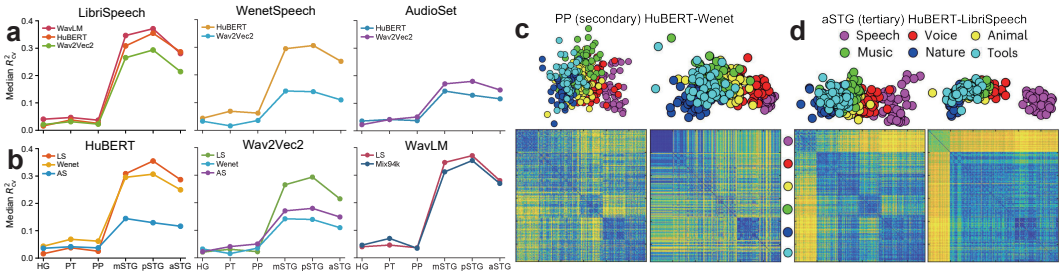

Figure 3: Learning-objective and pretraining-dataset associations with auditory cortical alignment. **a**, Median cross-validated cortical predictivity (R^2_{cv}) across the six auditory ROIs for self-supervised models pretrained on the same corpus but using different learning objectives, shown separately for LibriSpeech, WenetSpeech, and AudioSet. **b**, Cortical predictivity across pretraining corpora within the HuBERT, Wav2Vec2, and WavLM model families. **c**, Representative multidimensional scaling (MDS) and representational dissimilarity matrix (RDM) visualizations for secondary auditory cortex (PP) and HuBERT-Wenet. **d**, Corresponding visualizations for tertiary auditory cortex (aSTG) and HuBERT-LS. Panels **c–d** provide qualitative examples of the representational geometry accompanying the quantitative differences shown in **a–b**.

other training characteristics, these comparisons characterize dataset-associated differences in cortical alignment rather than isolating the causal contribution of any single property of the pretraining data.

4.3 Self-supervised cortical advantages do not extend to perceptual behavior

Given the strong tertiary-cortical alignment of several self-supervised models, we next asked whether these models also provided stronger accounts of human perceptual behavior (Fig. 4). Behavioral judgments were summarized as representational dissimilarity matrices (RDMs) separately for the sound- and word-judgment tasks. Representative geometries differed qualitatively between supervised YamNet and self-supervised HuBERT-AS, despite both using AudioSet-derived training data (Fig. 4a–c). YamNet showed clearer living–nonliving separation, whereas the categories overlapped more strongly in HuBERT-AS. We therefore quantified behavioral predictivity across all models using cross-validated R^2_{cv} (Fig. 4e).

In the sound-judgment task, supervised models provided the strongest DNN-based behavioral predictions. YamNet achieved the highest median R^2_{cv} among all evaluated models (0.306), followed by VGGish (0.229). The semantic GNewsW2V model (0.210) and the acoustic AudDims model (0.200) also showed higher median predictivity than the strongest self-supervised model, WavLM-Mix94k (0.198).

In the word-judgment task, semantic models showed the highest predictive performance, with median R^2_{cv} values ranging from 0.244 to 0.381. Among the DNN models, YamNet (0.156) and VGGish (0.117) again exceeded the strongest self-supervised model, WavLM-Mix94k (0.099), whereas Kell (0.085) did not. AudDims (0.089) likewise remained below WavLM-Mix94k. Relative to WavLM-Mix94k, YamNet's median R^2_{cv} was **54.4%** higher in the sound-judgment task and **56.8%** higher in the word-judgment task.

The MkNN analysis provided a complementary measure of local representational similarity (Fig. 4f). Across a broad range of neighborhood sizes, VGGish and YamNet showed greater neighborhood overlap with human behavioral representations than the self-supervised models, with the difference most apparent at intermediate neighborhood sizes. This local-topology result was consistent with the model ranking observed using predictive RSA.

Together, these results show that the self-supervised models exhibiting strong alignment with tertiary auditory cortex did not show a corresponding advantage in predicting human perceptual judgments. Thus, the model-class ranking observed for tertiary cortical prediction does not generalize to behavioral prediction, revealing a dissociation between cortical and behavioral alignment in the models evaluated here.

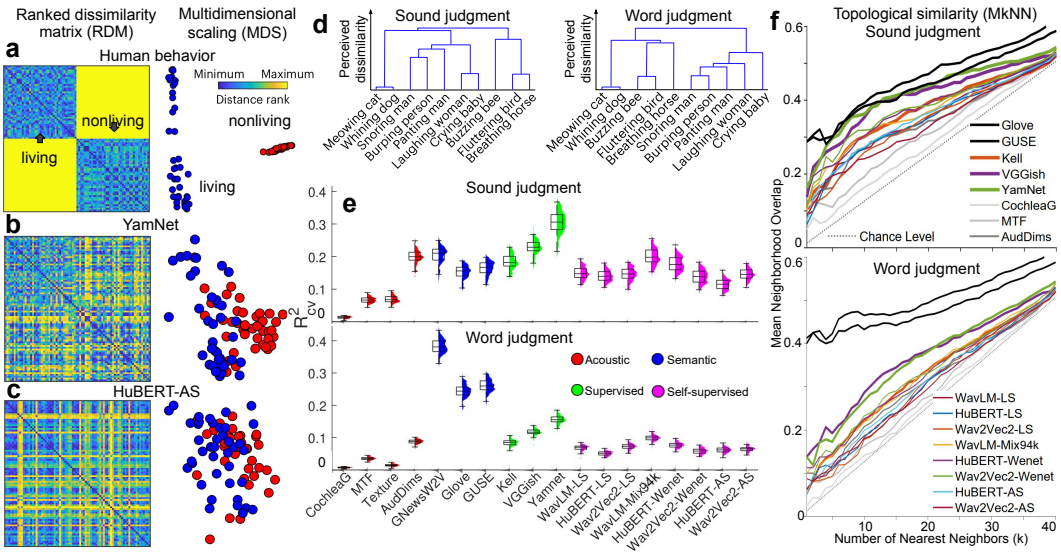

Figure 4: Behavioral representational geometry and model prediction of human perceptual judgments. **a**, Human behavioral RDM and multidimensional scaling (MDS) representation of living and nonliving stimuli. **b–c**, Representative RDM and MDS visualizations for supervised YamNet and self-supervised HuBERT-AS, respectively. **d**, Hierarchical dendrograms for the sound- and word-judgment tasks. **e**, Model-by-model cross-validated prediction (R^2_{cv}) across acoustic, semantic, supervised, and self-supervised models. Colored distributions show performance across cross-validation folds; box plots indicate the median, interquartile range, and whiskers extending to 1.5 times the interquartile range. **f**, Mutual k-nearest-neighbor (MkNN) overlap between model and behavioral representations across neighborhood sizes. $N = 20$ participants per task.

4.4 Supervised and self-supervised models show distinct layer-wise behavioral trajectories

We next asked whether the behavioral differences between supervised and self-supervised models were also reflected in where behaviorally predictive representations emerged across network depth (Fig. 5). We therefore evaluated behavioral predictivity separately at each extracted layer of the eleven DNN models.

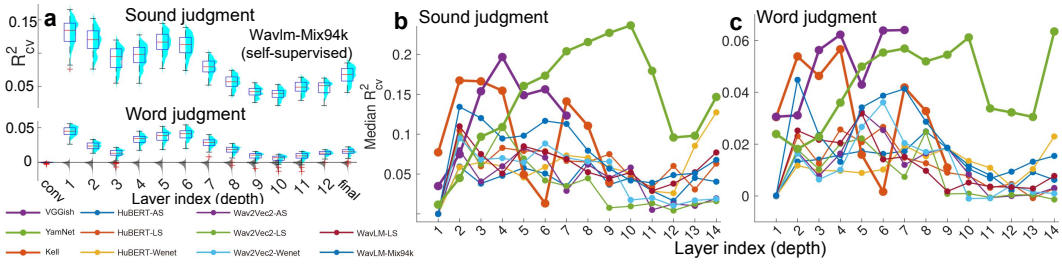

Figure 5: Layer-wise behavioral predictivity across supervised and self-supervised DNNs. **a**, Cross-validated predictive performance (R^2_{cv}) across layers of the self-supervised WavLM-Mix94k model for sound and word judgments. Colored distributions show performance across cross-validation folds, with overlaid box plots indicating the median, interquartile range, and whiskers extending to 1.5 times the interquartile range; gray distributions indicate the permutation baseline. **b–c**, Median R^2_{cv} across layers for all supervised and self-supervised DNNs in the sound-judgment (**b**) and word-judgment (**c**) tasks. Thick lines denote supervised models and thin lines denote self-supervised models. Layer indices indicate within-model depth and are not architecturally matched across models. $N = 20$ participants per task.

WavLM-Mix94k, the strongest self-supervised model in the overall behavioral comparison, provides a representative example (Fig. 5a). For both sound and word judgments, predictive performance was strongest at relatively early-to-intermediate network depths and generally declined toward deeper layers, although smaller secondary peaks were present at intermediate depths.

Across models, most self-supervised trajectories reached their maxima at early-to-intermediate depths in the sound-judgment task (Fig. 5b). HuBERT-Wenet was a notable exception, showing its strongest prediction at a relatively late layer. In contrast, the stronger supervised models tended to peak later: YamNet reached its maximum at a relatively late network stage, whereas VGGish peaked around intermediate depth. Kell was an exception among the supervised models, reaching its maximum substantially earlier.

A similar tendency was observed for word judgments (Fig. 5c). Most self-supervised models reached their maximum before or around the middle of the network, whereas VGGish and YamNet showed their strongest predictivity at later relative depths. This broad pattern was therefore observed across both behavioral tasks, despite model-specific exceptions. Together, these results indicate that behaviorally predictive representations tended to emerge earlier in the self-supervised models and later in the stronger supervised models evaluated here. However, exceptions in both model classes indicate that peak depth is not determined by supervision alone.

5 Discussion

Our results reveal a dissociation between cortical and behavioral alignment in auditory models. Several self-supervised models showed higher predictive alignment than supervised baselines in tertiary auditory cortex, whereas supervised models generally provided better predictions of human perceptual judgments. Across the models evaluated here, stronger cortical alignment did not necessarily imply stronger behavioral alignment. Instead, the relative advantage of a model depended on the biological target and cortical stage being modeled.

One possible interpretation of this dissociation is that supervised and self-supervised objectives emphasize different aspects of sensory representations. Supervised sound-classification models are optimized for task-relevant distinctions, which may contribute to their stronger behavioral alignment and later layer-wise behavioral prediction peaks. Self-supervised models instead optimize contrastive, masked-prediction, or denoising objectives and may preserve acoustic and contextual structure that aligns well with cortical population geometry without directly matching perceptual judgments. These observations are consistent with, but do not establish, a causal role of supervision or a specific top-down mechanism.

Cortical alignment also varied substantially within the self-supervised model class. The relative ordering of WavLM, HuBERT, and Wav2Vec2 differed across pretraining corpora, while changing the corpus within a model family altered the predictivity profile across the auditory hierarchy. Thus, neither a single learning objective nor a single pretraining corpus was uniformly optimal across cortical stages. These corpus-associated differences should not be interpreted causally, however, because LibriSpeech, WenetSpeech, AudioSet, and Mix94k differ simultaneously in language, duration, acoustic content, recording conditions, and other characteristics. Controlled experiments will be required to isolate which properties of the pretraining environment account for these differences.

Limitations Several limitations qualify these conclusions. The fMRI dataset included only five participants, and behavioral and fMRI responses were obtained from different participants, stimuli, and tasks, precluding direct within-subject tests of brain–behavior correspondence. Our benchmark does not cover all auditory self-supervised architectures, and the supervised and self-supervised models are not fully matched in architecture, training data, model scale, or training procedure. In addition, the pretraining corpora differ simultaneously in language, duration, acoustic content, recording conditions, and source domain, so the observed dataset-associated differences cannot isolate the causal contribution of any single corpus property. Finally, whether these findings generalize to other auditory tasks, neural measurement modalities, datasets, or sensory systems remains open.

AI USE STATEMENT

Generative AI tools were used only for debugging, LaTeX formatting, and language polishing of the manuscript. All AI-assisted content was carefully reviewed and verified by the authors, who take full responsibility for the accuracy and integrity of the final manuscript.

ETHICS STATEMENT

This study did not recruit new human participants or collect new human-subject data. We analyzed previously collected and publicly available behavioral and fMRI datasets released by the original studies. The behavioral dataset was originally collected under protocols approved by the McGill Research Ethics Board, and the fMRI dataset was originally collected under the corresponding ethical procedures approved by Maastricht University. Our study involved only secondary analysis of these existing datasets.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the computational models, datasets, feature extraction procedures, representational analyses, cross-validation schemes, statistical inference procedures, and evaluation metrics in the Methods and Appendix. Model identities and pretrained checkpoints are specified in the Appendix, and the behavioral and fMRI datasets used in this study are publicly available from the sources cited in the paper. Code and supporting materials for the principal analyses are provided as anonymized supplementary material. The supplementary package includes core analysis code, derived representational inputs, and supporting result files associated with the reported analyses.

REFERENCES

Bilal Ahmed, Joshua D. Downer, Brian J. Malone, and Joseph G. Makin. Deep neural networks explain spiking activity in auditory cortex. *PLOS Computational Biology*, 21(8):e1013334, 2025.

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

Thierry Bertin-Mahieux, Daniel P. W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011)*, 2011.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

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

Taishih Chi, Powen Ru, and Shihab A. Shamma. Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, 118(2):887–906, 2005.

Alain De Cheveigné and Hideki Kawahara. YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930, 2002.

Andrew Francl and Josh H. McDermott. Deep neural network models of sound localization reveal how perception is adapted to real-world environments. *Nature Human Behaviour*, 6(1):111–133, 2022.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. AudioSet: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780. IEEE, 2017.

Bruno L. Giordano, John McDonnell, and Stephen McAdams. Hearing living symbols and nonliving icons: Category specificities in the cognitive processing of environmental sounds. *Brain and Cognition*, 73(1):7–19, 2010.

Bruno L. Giordano, Michele Esposito, Giancarlo Valente, and Elia Formisano. Intermediate acoustic-to-semantic representations link behavioral and neural responses to natural sounds. *Nature Neuroscience*, 26(4):664–672, 2023.

Ian M. Griffith, R. Preston Hess, and Josh H. McDermott. Optimized feature gains explain and predict successes and failures of human selective listening. *Nature Human Behaviour*, pp. 1–23, 2026.

Umut Güçlü, Jordy Thielen, Michael Hanke, and Marcel Van Gerven. Brains on beats. *Advances in Neural Information Processing Systems*, 29, 2016.

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.

Alexander J. E. Kell, Daniel L. K. Yamins, Erica N. Shook, Sam V. Norman-Haignere, and Josh H. McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644, 2018.

Menoua Keshishian, Hassan Akbari, Bahar Khalighinejad, Jose L. Herrero, Ashesh D. Mehta, and Nima Mesgarani. Estimating and interpreting nonlinear receptive field of sensory neural responses with deep neural network models. *eLife*, 9:e53445, 2020.

Yuanning Li, Gopala K. Anumanchipalli, Abdelrahman Mohamed, Peili Chen, Laurel H. Carney, Junfeng Lu, Jinsong Wu, and Edward F. Chang. Dissecting neural computations in the human auditory pathway using deep neural networks for speech. *Nature Neuroscience*, 26(12):2213–2225, 2023.

Josh H. McDermott and Eero P. Simoncelli. Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis. *Neuron*, 71(5):926–940, 2011.

Johannes Mehrer, Courtney J. Spoerer, Emer C. Jones, Nikolaus Kriegeskorte, and Tim C. Kietzmann. An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, 118(8):e2011417118, 2021.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Juliette Millet and Jean-Remi King. Inductive biases, pretraining and fine-tuning jointly account for brain responses to speech. *arXiv preprint arXiv:2103.01032*, 2021.

Juliette Millet, Charlotte Caucheteux, Yves Boubenec, Alexandre Gramfort, Ewan Dunbar, Christophe Pallier, and Jean-Remi King. Toward a realistic model of speech processing in the brain with self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 33428–33443, 2022.

Gavin Mischler, Menoua Keshishian, Stephan Bickel, Ashesh D. Mehta, and Nima Mesgarani. Deep neural networks effectively model neural adaptation to changing background noise and suggest nonlinear noise filtering methods in auditory cortex. *NeuroImage*, 266:119819, 2023.

Sam V. Norman-Haignere, Menoua Keshishian, Orrin Devinsky, Werner Doyle, Guy M. McKhann, Catherine A. Schevon, Adeen Flinker, and Nima Mesgarani. Temporal integration in human auditory cortex is predominantly yoked to absolute time. *Nature Neuroscience*, 28(11):2356–2365, 2025.

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.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

Kyle M. Rupp, Jasmine L. Hect, Emily E. Harford, Lori L. Holt, Avniel Singh Ghuman, and Taylor J. Abel. A hierarchy of processing complexity and timescales for natural sounds in the human auditory cortex. *Proceedings of the National Academy of Sciences*, 122(18):e2412243122, 2025.

Mark R. Saddler and Josh H. McDermott. Models optimized for real-world tasks reveal the task-dependent necessity of precise temporal coding in hearing. *Nature Communications*, 15(1):10590, 2024.

Mark R. Saddler, Ray Gonzalez, and Josh H. McDermott. Deep neural network models reveal interplay of peripheral coding and stimulus statistics in pitch perception. *Nature Communications*, 12(1):7278, 2021.

Roberta Santoro, Michelle Moerel, Federico De Martino, Giancarlo Valente, Kamil Ugurbil, Essa Yacoub, and Elia Formisano. Reconstructing the spectrotemporal modulations of real-life sounds from fMRI response patterns. *Proceedings of the National Academy of Sciences*, 114(18):4799–4804, 2017.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-Score: Which artificial neural network for object recognition is most brain-like? *bioRxiv preprint*, 2018. URL https://www.biorxiv.org/content/10.1101/407007v2.

Greta Tuckute, Jenelle Feather, Dana Boebinger, and Josh H. McDermott. Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions. *PLOS Biology*, 21(12):e3002366, 2023.

Aditya R. Vaidya, Shailee Jain, and Alexander G. Huth. Self-supervised models of audio effectively explain human cortical responses to speech. *arXiv preprint arXiv:2205.14252*, 2022.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 993–1003, 2021.

Daniel L. K. Yamins and James J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, 2016.

Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, et al. WenetSpeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. In *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6182–6186. IEEE, 2022.

A Computational models

A.1 Acoustic models

To characterize acoustic and perceptually relevant properties of the sound stimuli, we evaluated four deterministic computational models based on biophysically or psychophysically motivated signal transformations: CochleaG, MTF, Texture, and AudDims. Unlike the supervised and self-supervised DNNs described below, these models do not require a pretraining dataset.

A.1.1 COCHLEAGRAM (COCHLEAG)

To characterize peripheral auditory filtering, we used a biologically inspired cochleagram representation implemented with NSLtools (Chi et al., 2005). The model transforms each acoustic stimulus into a time-varying spectrotemporal representation using 128 log-spaced frequency filters spanning 179–7,246 Hz at an 8 ms temporal resolution. From this representation, we extracted two components for subsequent analyses: the full time-varying cochleagram and a time-averaged 128-dimensional cochleagram summarizing spectral energy across the duration of the sound.

A.1.2 MODULATION TRANSFER FUNCTION (MTF)

To characterize spectrotemporal modulations, we applied a multiresolution spectrotemporal model implemented with NSLtools (Chi et al., 2005) to the cochleagram representation. The model analyzes acoustic activity across spectral scales from 0.25 to 8 cycles/octave and temporal modulation rates from 4 to 156 Hz at an 8 ms temporal resolution, yielding complex-valued spectrotemporal representations.

We extracted six MTF components: three time-varying representations (magnitude, phase, and combined magnitude/phase) and their corresponding three time-averaged representations. The time-varying components contained up to 9,856,000 feature values for the behavioral stimuli and 4,928,000 for the fMRI stimuli, whereas the time-averaged components contained 19,712 features for both datasets.

A.1.3 TEXTURE MODEL (TEXTURE)

To characterize higher-order acoustic statistics of sound textures, we used the sound texture model developed by McDermott & Simoncelli (2011) (Sound Texture Synthesis Toolbox v1.7). The model computes summary statistics from the time-varying amplitude envelopes of a cochleagram containing 32 frequency bands spaced on an equivalent rectangular bandwidth (ERB)-rate scale from 20 Hz to 10 kHz, together with modulation analyses using six filters log-spaced between 30 and 100 Hz.

We extracted five statistical components: marginal envelope statistics (mean, variance, skewness, and kurtosis), pairwise envelope correlations, modulation-band power, constant-modulation-frequency correlations (C1), and adjacent-modulation-frequency correlations (C2). Together, these components yielded a 6,656-dimensional representation for each stimulus.

A.1.4 AUDITORY DIMENSIONS (AUDDIMS)

To characterize perceptually relevant acoustic properties, we modeled five time-varying auditory dimensions at a 1 ms temporal resolution: pitch, loudness, periodicity, timbre brightness, and roughness. Time-varying loudness and timbre brightness, defined as spectral centroid, were derived from instantaneous specific-loudness estimates using the Genesis Loudness Toolbox. Pitch and periodicity, with periodicity defined as the ratio of periodic to aperiodic power, were estimated using the YIN algorithm (De Cheveigné & Kawahara, 2002), while roughness was computed using MIRtoolbox.

For each auditory dimension, we extracted three components: the complete time-varying profile, its temporal mean, and its temporal standard deviation. This yielded 15 components in total, corresponding to 50,005 feature values for the behavioral stimuli and 10,005 for the fMRI stimuli.

A.2 SEMANTIC MODELS

To characterize the semantic properties associated with the sound stimuli, we used three pretrained natural language processing (NLP) models to generate embeddings from the corresponding descriptive verb-and-noun labels.

First, we used GloVe (Pennington et al., 2014), a distributional word-embedding model trained on large text corpora including Wikipedia, Common Crawl, and Gigaword. GloVe represents each word as a 300-dimensional vector. Second, we used GNewsW2V (Mikolov et al., 2013), a Word2Vec model pretrained on approximately 100 billion words from Google News, which also produces 300-dimensional word embeddings. For both GloVe and GNewsW2V, the semantic representation of each sound stimulus was obtained by averaging the embeddings of its corresponding verb and noun.

Finally, we used the Google Universal Sentence Encoder (GUSE) (Cer et al., 2018), which produces a 512-dimensional representation and was pretrained on diverse web-based text sources, including Wikipedia, news articles, question-answer pages, and discussion forums. Because GUSE is designed to encode multiword text inputs, the combined verb-and-noun description of each sound was provided directly as a single input rather than averaging separate word embeddings.

Each of the three semantic models contributed one representation component to the subsequent predictive analyses.

A.3 SUPERVISED DNN MODELS

We evaluated three pretrained supervised auditory DNNs: Kell's dual-task network, VGGish, and YamNet. All three operate on two-dimensional time-frequency representations rather than directly on raw audio waveforms, using either a biologically inspired cochleagram or a stabilized log-mel spectrogram as input. For each model, activations were extracted from prespecified intermediate and higher-level layers and treated as separate model components in the predictive analyses.

A.3.1 KELL'S DUAL-TASK NETWORK (KELL)

Description. The Kell network processes 2-second audio excerpts represented as two-dimensional cochleagrams. Its architecture contains a shared trunk of convolutional blocks followed by task-specific branches for two supervised auditory classification objectives. Each branch contains additional convolutional, pooling, and fully connected layers before the final classification output.

Model outputs. We extracted activations from nine prespecified model components. Across these components, the extracted representations contained a total of 557,152 feature values per stimulus, with individual components ranging from 1,024 to 177,504 features.

Pretraining dataset and objective. The network was jointly optimized for two auditory classification tasks: 587-way word recognition using speech drawn from the TIMIT and Wall Street Journal corpora, and 41-way music-genre recognition using the Million Song Dataset (Bertin-Mahieux et al., 2011).

A.3.2 VGGISH

Description. VGGish is a VGG-style convolutional network adapted for audio. It processes 0.975-second audio windows represented as stabilized log-mel spectrograms. The architecture contains four convolutional blocks followed by three fully connected layers that progressively reduce the dimensionality of the representation.

Model outputs. We extracted activations from seven prespecified model components. The final fully connected representation is 128-dimensional, and the extracted components together contained 192,640 feature values per stimulus. For behavioral stimuli longer than a single 0.975-second input window, activations were extracted from consecutive windows and averaged across time.

Pretraining dataset and objective. VGGish was pretrained on a large collection of 10-second YouTube audio segments using supervised audio-event classification targets defined by the AudioSet ontology.

A.3.3 YAMNET

Description. YamNet is an efficient supervised audio-classification network based on the MobileNetV1 architecture. Like VGGish, it operates on 0.975-second stabilized log-mel spectrograms. The network begins with a standard convolutional block followed by 13 depthwise-separable convolutional blocks composed of depthwise convolution, batch normalization, ReLU activation, pointwise convolution, batch normalization, and ReLU activation.

Model outputs. The final convolutional representation contains 1,024 channels and is globally averaged to produce a 1,024-dimensional embedding before the final classification layer. We extracted

activations from 14 prespecified model components, corresponding to a total of 380,928 feature values per stimulus.

Pretraining dataset and objective. YamNet was pretrained on 10-second YouTube audio segments to classify 521 sound-event categories from the AudioSet ontology.

A.4 Self-supervised DNN models

A.4.1 Shared architecture and feature extraction

To investigate the representational geometry of self-supervised speech models, we evaluated eight pretrained networks from three model families: Wav2Vec2 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), and WavLM (Chen et al., 2022). We restricted the analysis to base-scale variants to obtain closely matched architectures across the evaluated models. All checkpoints were accessed through the Hugging Face Transformers library.

The evaluated models share a common two-stage organization: a convolutional feature encoder operating directly on 16-kHz raw audio, followed by a Transformer-based contextual network. Although the model families are closely related architecturally, they are not strictly identical; in particular, WavLM differs from Wav2Vec2 and HuBERT in aspects of its positional representation and self-attention implementation.

Local feature encoder. The convolutional front end consists of seven temporal convolutional layers with 512 channels. The convolutional kernel sizes are $(10, 3, 3, 3, 3, 2, 2)$, with strides $(5, 2, 2, 2, 2, 2, 2)$, yielding a total stride of 320 audio samples. For 16-kHz input audio, this corresponds to one latent feature vector every 20 ms, with an effective receptive field of approximately 25 ms.

Contextual representations. The convolutional representations are subsequently processed by a stack of 12 Transformer encoder layers with a hidden-state dimensionality of 768. Wav2Vec2 and HuBERT use a convolution-based positional embedding before the Transformer stack, whereas WavLM incorporates a gated relative-position mechanism within its attention architecture. These differences are one reason we treat the evaluated model families as closely matched rather than architecturally identical.

Activation extraction. For downstream analyses, raw audio waveforms were processed using the corresponding Hugging Face input preprocessing pipeline. Prespecified representations from the convolutional encoder and successive contextual layers were extracted and treated as model components in the predictive RSA analyses. Layer-wise comparisons therefore reflect within-model depth and should not be interpreted as exact architectural correspondence between different model families.

A.4.2 Self-supervised pretraining objectives

Although Wav2Vec2, HuBERT, and WavLM share a closely related base-scale architecture, their self-supervised pretraining objectives differ in how prediction targets are constructed and how masked audio representations are learned.

Wav2Vec2: contrastive learning with online quantization. Wav2Vec2 uses a contrastive objective with online target generation. During pretraining, portions of the latent representations produced by the convolutional encoder are masked before being passed to the contextual network. In parallel, latent encoder outputs are quantized using a Gumbel–Softmax-based module to construct discrete target representations. At masked time steps, the contextual representation is trained to distinguish the corresponding quantized target from a set of distractor representations. Thus, both the target representation and contextual network are learned during pretraining.

HuBERT: masked prediction with offline clustering. HuBERT replaces the online contrastive targets of Wav2Vec2 with discrete pseudo-labels generated by offline clustering. Acoustic features, such as MFCCs or representations obtained from an earlier HuBERT iteration, are clustered using

k-means to assign a discrete hidden-unit label to each frame. During pretraining, portions of the continuous input representation are masked and the Transformer is trained to predict these discrete targets. This objective therefore emphasizes contextual prediction of externally defined hidden-unit targets rather than contrastive discrimination among online-quantized representations.

WavLM: masked prediction with denoising. WavLM extends masked hidden-unit prediction by introducing additional denoising during pretraining. As in HuBERT, discrete pseudo-labels are generated using offline clustering. During training, however, input waveforms can be corrupted by background noise or mixed with speech from additional speakers, while the model is trained to predict the targets associated with the underlying speech. This procedure introduces an additional training pressure for representations that remain informative under acoustic interference, while retaining the masked-prediction framework.

A.4.3 PRETRAINING CORPORA

The evaluated checkpoints were pretrained on four auditory corpora: LibriSpeech, Mix94k, WenetSpeech, and AudioSet. These corpora differ in language, duration, acoustic content, recording conditions, and source domain. Accordingly, the comparisons in the main text characterize corpus-associated differences and should not be interpreted as isolating the causal contribution of any single dataset property.

LibriSpeech. LibriSpeech is a 960-hour corpus of read English speech derived from LibriVox audiobooks (Panayotov et al., 2015). The corpus comprises the commonly used 100-hour and 360-hour "clean" subsets together with a 500-hour "other" subset. In the present benchmark, LibriSpeech provides a common pretraining corpus for WavLM-LS, HuBERT-LS, and Wav2Vec2-LS, allowing comparison of different self-supervised objectives under a shared speech corpus.

Mix94k. Mix94k is the approximately 94,000-hour composite speech corpus used to pretrain WavLM Base+ (Chen et al., 2022). It combines approximately 60,000 hours of Libri-Light, 10,000 hours of GigaSpeech, and 24,000 hours of English VoxPopuli speech (Wang et al., 2021). Relative to LibriSpeech, this corpus is substantially larger and combines speech from multiple source domains.

WenetSpeech. WenetSpeech is a large-scale Mandarin speech corpus collected from internet audio sources such as YouTube and podcasts (Zhang et al., 2022). The evaluated HuBERT-Wenet and Wav2Vec2-Wenet checkpoints use approximately 10,000 hours of high-confidence speech drawn from this corpus. The data span multiple content domains and include broader variation in speaking style and recording conditions than read-audiobook speech.

AudioSet. AudioSet is a large-scale collection of 10-second YouTube audio clips covering a broad range of everyday sound events (Gemmeke et al., 2017). The full dataset contains more than 2.1 million clips, corresponding to approximately 5,800 hours of audio, and includes human and animal vocalizations, music, environmental sounds, and mechanical events. In contrast to the speech-focused corpora above, AudioSet contains heterogeneous nonspeech as well as speech-related acoustic content.

A.4.4 EVALUATED SELF-SUPERVISED PRETRAINED MODELS

For reproducibility, we report the exact pretrained checkpoints used in our analyses. All checkpoints were obtained through the Hugging Face Transformers library.

LibriSpeech (English read speech): WavLM-LS (`microsoft/wavlm-base`), HuBERT-LS (`facebook/hubert-base-ls960`), and Wav2Vec2-LS (`facebook/wav2vec2-base`).

Mix94k (large-scale multi-domain speech): WavLM-Mix94k (`microsoft/wavlm-base-plus`).

WenetSpeech (Mandarin multi-domain speech): HuBERT-Wenet (`TencentGameMate/chinese-hubert-base`) and Wav2Vec2-Wenet (`TencentGameMate/chinese-wav2vec2-base`).

AudioSet (general sound events): HuBERT-AS (`ALM/hubert-base-audioset`) and Wav2Vec2-AS (`ALM/wav2vec2-base-audioset`).

For Wav2Vec2-LS, we used the pretrained `facebook/wav2vec2-base` checkpoint rather than an ASR-fine-tuned variant. The analyzed representations were therefore obtained prior to supervised task-specific fine-tuning.

B Experiments and model representation analyses

B.1 Behavioral and high-field 7T fMRI data

We analyzed previously collected behavioral data from Giordano et al. (2010), obtained from the Dryad repository (`https://datadryad.org/dataset/doi:10.5061/dryad.0p2ngf258`). The original behavioral study was conducted under protocols approved by the McGill Research Ethics Board. Forty participants were divided into two groups of 20 and completed hierarchical similarity judgments involving 80 natural sounds, comprising 40 living and 40 nonliving sound sources. One group performed judgments using the sounds themselves, whereas the other used corresponding verb-and-noun descriptions (e.g., "crying baby").

Participants iteratively grouped stimuli according to perceived similarity until all items were merged into a single hierarchical structure. Living and nonliving stimuli were judged separately. Pairwise behavioral dissimilarities were derived from the level of the resulting hierarchy at which each pair of stimuli was merged, yielding separate behavioral representations for the sound- and word-judgment tasks.

High-field 7T fMRI responses were obtained from Santoro et al. (2017) through the same Dryad repository. The original fMRI experiment was conducted under protocols approved by Maastricht University. Five participants listened to 288 one-second natural sounds spanning six categories: speech, voice, animal, music, nature, and tools. Stimuli were presented using an event-related design during a 1.4-s silent interval between volume acquisitions to reduce interference from scanner noise. Participants performed an incidental one-back repetition-detection task; repetition trials were excluded from the analyses.

Voxel-specific hemodynamic response functions were estimated in the original analysis pipeline, and stimulus-specific fMRI responses were represented by general linear model (GLM) coefficients. Responses were analyzed within six bilateral auditory regions of interest (ROIs): Heschl's gyrus (HG), planum temporale (PT), planum polare (PP), middle superior temporal gyrus (mSTG), posterior superior temporal gyrus (pSTG), and anterior superior temporal gyrus (aSTG). Following the organization used in the main text, these ROIs were grouped into primary (HG), secondary (PT and PP), and tertiary (mSTG, pSTG, and aSTG) auditory cortex.

B.2 Cross-validation and predictive representational analysis

We evaluated alignment between computational models and human behavioral and fMRI data using the predictive representational similarity analysis (RSA) framework described in the main text (Fig. 1). Behavioral representational dissimilarity matrices (RDMs) were derived from the hierarchical similarity judgments described above. For the fMRI data, ROI-specific RDMs were constructed separately for each participant by computing Euclidean distances between demeaned stimulus-wise GLM response patterns. For each computational model component, model RDMs were constructed using pairwise cosine distances between stimulus representations.

Predictive alignment was evaluated using cross-validated ordinary least-squares regression. Within each training fold, model-derived representational distances were used to predict group-averaged human representational distances. The fitted regression coefficients were then held fixed and evaluated on the corresponding held-out human data.

Cross-validation was adapted to the experimental structure of each dataset. For the behavioral analyses, we used 100 random participant split-halves within each task, with 10 participants assigned to the training set and 10 to the test set in each split. Because the living and nonliving stimulus sets were evaluated separately and did not form a common cross-category representational space, behavioral cross-validation was performed across participants rather than across stimuli.

For the fMRI analyses, the 288 stimuli were organized into four non-overlapping, categorically matched sets of 72 stimuli. We evaluated all 10 exhaustive participant splits, using three participants for training and two for testing. Within each participant split, three stimulus sets were used for fitting the predictive model and the remaining set was held out for evaluation.

Predictive performance was quantified using the cross-validated coefficient of determination,

$$R^2_{\mathrm{cv}} = 1 - \frac{SSE_{\mathrm{test}}}{SST_{\mathrm{test}}}.$$

Statistical significance was assessed using 10,000 matrix permutations. Inference was performed on the median R^2_{cv} across cross-validation splits, with a maximum-statistic procedure used to control the family-wise error rate at FWER ≤ 0.05.

B.2.1 Cross-dataset fMRI-to-behavior transfer analysis

As an auxiliary analysis, we also examined whether model components that predicted cortical representational geometry could transfer to the separate behavioral stimulus set. Within each fMRI cross-validation fold, regression coefficients were estimated to predict ROI-specific fMRI representational distances from DNN-derived representational distances. These fitted coefficients were then applied to the corresponding DNN representations of the behavioral stimuli, producing an fMRI-weighted prediction of behavioral representational geometry.

Predictions were averaged across fMRI cross-validation folds to obtain one fMRI-grounded behavioral predictor for each ROI. These predictors were then evaluated against the observed behavioral dissimilarities using the same participant-based cross-validation and permutation framework described above. We evaluated individual ROIs as well as the combined ROI representation. We additionally applied commonality analysis to characterize the unique and shared behavioral variance associated with predictors derived from HG and pSTG.

This transfer analysis was intended as a complementary test of whether model-derived representational components associated with cortical responses also captured variance in perceptual judgments across a distinct stimulus set; it was not used as the primary measure of either cortical or behavioral alignment.

B.3 Derivation of pairwise dimensionalities

Predictive RSA operates on the unique pairwise distances contained in each RDM. For a representation containing n stimuli, the number of unique off-diagonal stimulus pairs is

$$\binom{n}{2} = \frac{n(n-1)}{2}.$$

Behavioral data. Living and nonliving stimuli were evaluated as separate sets of 40 stimuli. Each 40-stimulus set therefore contained

$$\binom{40}{2} = \frac{40 \times 39}{2} = 780$$

unique pairwise distances. Thus, each behavioral RDM contributed 780 distance observations to the predictive analysis.

fMRI data. The fMRI experiment contained 288 stimuli divided into four non-overlapping sets of 72 stimuli. In each stimulus cross-validation fold, three sets (216 stimuli) were used for model fitting and the remaining set (72 stimuli) was held out for evaluation. The training RDM therefore contained

$$\binom{216}{2} = \frac{216 \times 215}{2} = 23{,}220$$

unique distances, whereas the held-out test RDM contained

$$\binom{72}{2} = \frac{72 \times 71}{2} = 2{,}556$$

unique distances. Accordingly, each fMRI stimulus cross-validation fold optimized the regression coefficients using 23,220 pairwise distances and evaluated them using 2,556 held-out distances.

B.4 Mutual k-nearest-neighbor analysis

Let $D^{(A)}$ and $D^{(B)}$ denote the representational dissimilarity matrices of a biological representation and a computational model, respectively, across N stimuli. For each stimulus i, let $\mathcal{N}_k^{(A)}(i)$ and $\mathcal{N}_k^{(B)}(i)$ denote the sets of its k nearest neighbors in the two representational spaces, excluding the stimulus itself.

The mutual k-nearest-neighbor (MkNN) overlap was defined as

$$\mathrm{MkNN}(k) = \frac{1}{N} \sum_{i=1}^{N} \frac{\left| \mathcal{N}_k^{(A)}(i) \cap \mathcal{N}_k^{(B)}(i) \right|}{k}. \tag{4}$$

MkNN was evaluated across neighborhood sizes k to characterize representational similarity at different local scales. For two independently organized representations, the expected chance-level overlap is $k/(N-1)$.

B.5 Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI)

As a supplementary characterization of category structure in DNN representations, we used the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) to compare model-derived cluster assignments with the reference stimulus-category labels.

The Rand Index quantifies agreement between two partitions by considering whether pairs of samples are assigned to the same or different clusters in each partition. ARI corrects the Rand Index for agreement expected by chance. An ARI of 1 indicates identical partitions, whereas values near zero indicate agreement close to the chance expectation; negative values are possible when agreement is below chance.

NMI quantifies the mutual information between two cluster assignments and normalizes it by their entropies. Under the normalization used here, NMI ranges from 0 to 1, with larger values indicating greater correspondence between the model-derived clustering and the reference category structure. Unlike ARI, NMI is normalized but is not explicitly corrected for chance.

C Supplementary Results

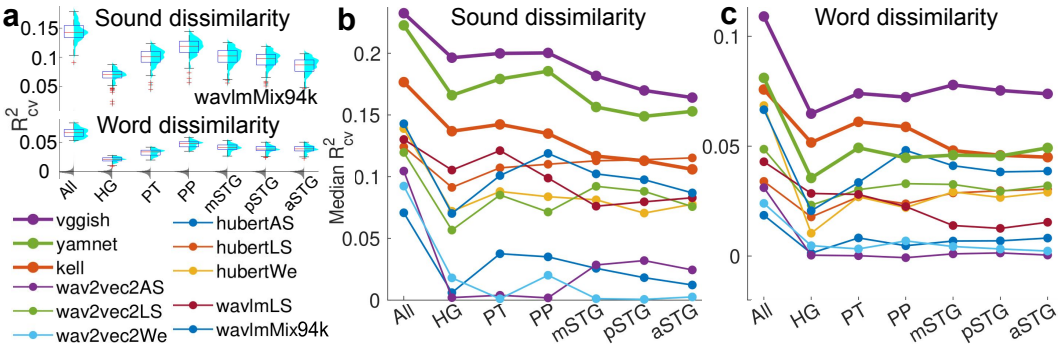

Figure 6: Cross-dataset prediction of behavioral representations from DNN-weighted fMRI representations. **a**, Cross-validated predictive performance (R^2_{cv}) across layers of WavLM-Mix94k for the sound- and word-judgment tasks. Colored distributions show performance across cross-validation folds, with overlaid box plots indicating the median, interquartile range, and whiskers extending to 1.5 times the interquartile range; gray distributions indicate the permutation baseline. **b**, Median R^2_{cv} for the combined ROIs and each individual auditory ROI across the three supervised and eight self-supervised DNNs in the sound-judgment task. **c**, Corresponding results for the word-judgment task. Thick lines denote supervised models and thin lines denote self-supervised models.

Table 2: Median cross-validated predictive performance (R^2_{cv}) for each model on the sound- and word-judgment tasks shown in Fig. 4e. The benchmark includes four acoustic, three semantic, three supervised DNN, and eight self-supervised DNN models.

Model	**Sound judgment**	**Word judgment**
CochleaG	0.0139	0.0077
MTF	0.0665	0.0356
Texture	0.0673	0.0144
AudDims	0.1995	0.0885
GNewsW2V	0.2104	0.3806
GloVe	0.1559	0.2438
GUSE	0.1670	0.2598
Kell	0.1817	0.0849
VGGish	0.2287	0.1168
YamNet	0.3058	0.1559
WavLM-LS	0.1476	0.0703
HuBERT-LS	0.1395	0.0513
Wav2Vec2-LS	0.1466	0.0731
WavLM-Mix94k	0.1981	0.0994
HuBERT-Wenet	0.1757	0.0772
Wav2Vec2-Wenet	0.1389	0.0586
HuBERT-AS	0.1153	0.0622
Wav2Vec2-AS	0.1448	0.0653

Table 3: Median cross-validated predictive performance (R^2_{cv}) for each model across the six auditory fMRI regions of interest (ROIs) shown in Fig. 2b–g. The benchmark includes four acoustic, three semantic, three supervised DNN, and eight self-supervised DNN models. ROIs are grouped as primary auditory cortex (HG: Heschl's gyrus), secondary auditory cortex (PT: planum temporale; PP: planum polare), and tertiary auditory cortex (mSTG, pSTG, and aSTG: middle, posterior, and anterior superior temporal gyrus, respectively). $N = 5$ fMRI participants.

Model	**HG**	**PT**	**PP**	**mSTG**	**pSTG**	**aSTG**
CochleaG	0.0324	0.0088	0.0126	0.0065	0.0070	0.0074
MTF	0.0625	0.0155	0.0211	0.0175	0.0186	0.0086
Texture	0.0227	0.0104	0.0113	0.0089	0.0137	0.0098
AudDims	0.0449	0.0269	0.0317	0.1057	0.1140	0.0804
GNewsW2V	0.0210	0.0451	0.0555	0.1288	0.1269	0.0979
GloVe	0.0127	0.0286	0.0351	0.0771	0.0748	0.0573
GUSE	0.0168	0.0275	0.0368	0.0691	0.0620	0.0487
Kell	0.0459	0.0459	0.0554	0.1474	0.1673	0.1258
VGGish	0.0614	0.0933	0.1110	0.2764	0.2931	0.2207
YamNet	0.0719	0.0926	0.1075	0.2319	0.2342	0.1571
WavLM-LS	0.0404	0.0466	0.0368	0.3464	0.3712	0.2806
HuBERT-LS	0.0159	0.0377	0.0250	0.3084	0.3548	0.2866
Wav2Vec2-LS	0.0209	0.0314	0.0225	0.2653	0.2934	0.2143
WavLM-Mix94k	0.0464	0.0706	0.0350	0.3123	0.3533	0.2701
HuBERT-Wenet	0.0433	0.0687	0.0618	0.2953	0.3067	0.2495
Wav2Vec2-Wenet	0.0325	0.0157	0.0353	0.1419	0.1395	0.1097
HuBERT-AS	0.0358	0.0414	0.0372	0.1445	0.1293	0.1164
Wav2Vec2-AS	0.0226	0.0410	0.0506	0.1707	0.1798	0.1485

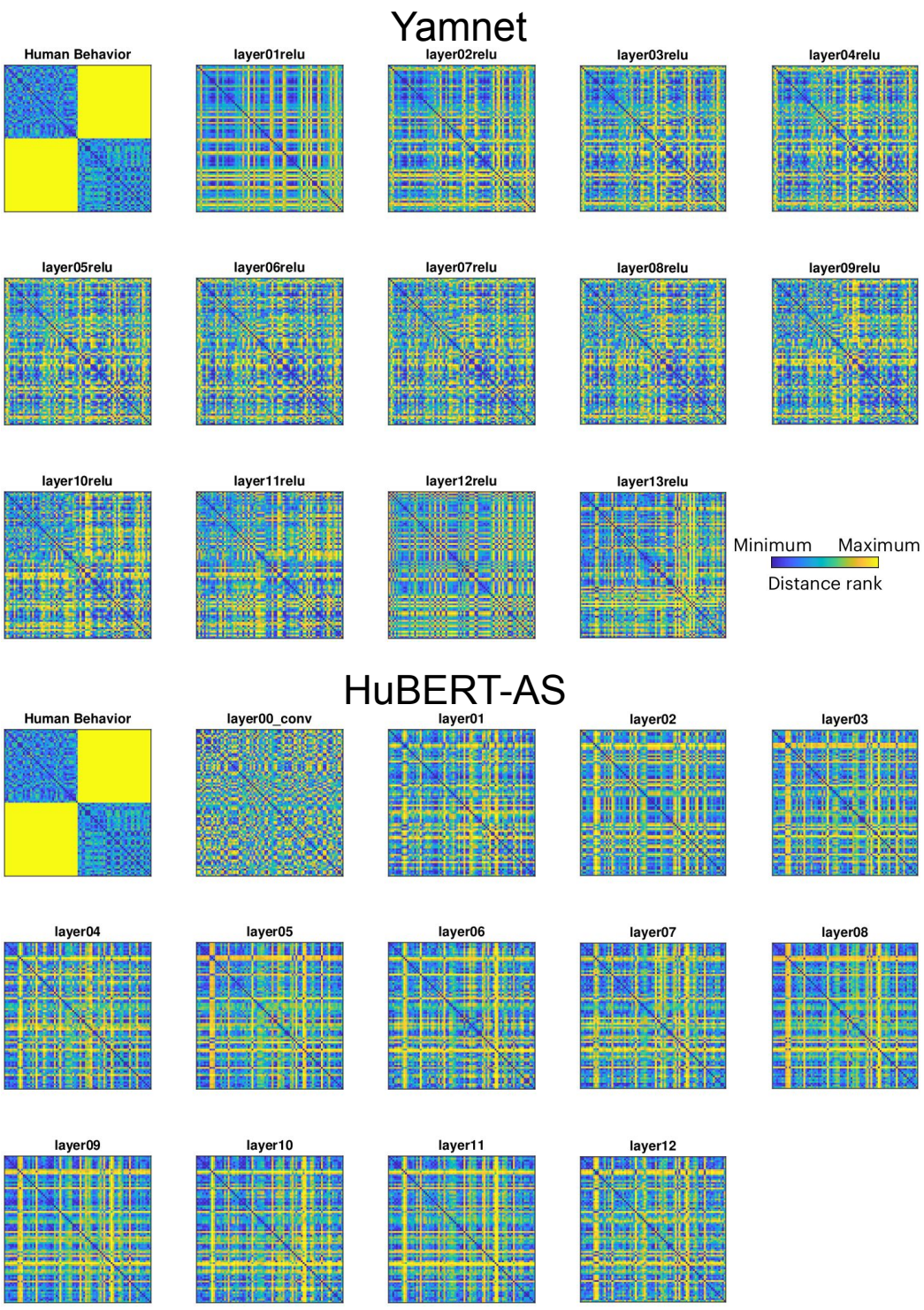

Figure 7: Layer-wise representational dissimilarity matrices (RDMs) for human behavior, supervised YamNet, and self-supervised HuBERT-AS. Both DNNs were pretrained using AudioSet-derived supervision or pretraining data, respectively. The highlighted model RDMs correspond to the representative behavioral geometries shown in Fig. 4b–c. For YamNet, the highlighted representation is `layer10relu`; for HuBERT-AS, it is `layer09`.

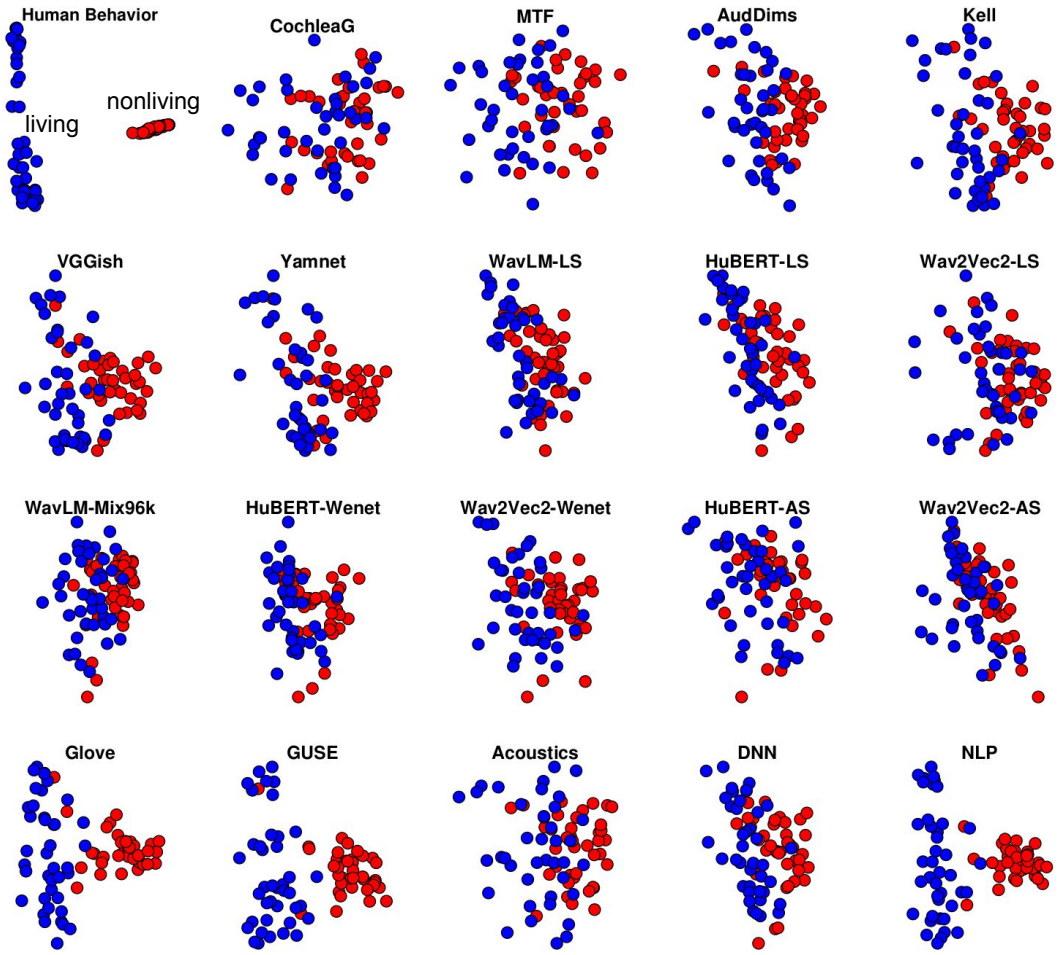

Figure 8: Metric multidimensional scaling (MDS) visualizations of natural-sound representations derived from human behavioral judgments, acoustic models, semantic models, supervised DNNs, self-supervised DNNs, and combined model representations. The YamNet and HuBERT-AS visualizations correspond to the representative model geometries shown in Fig. 4b–c. Colors denote living and nonliving stimuli.

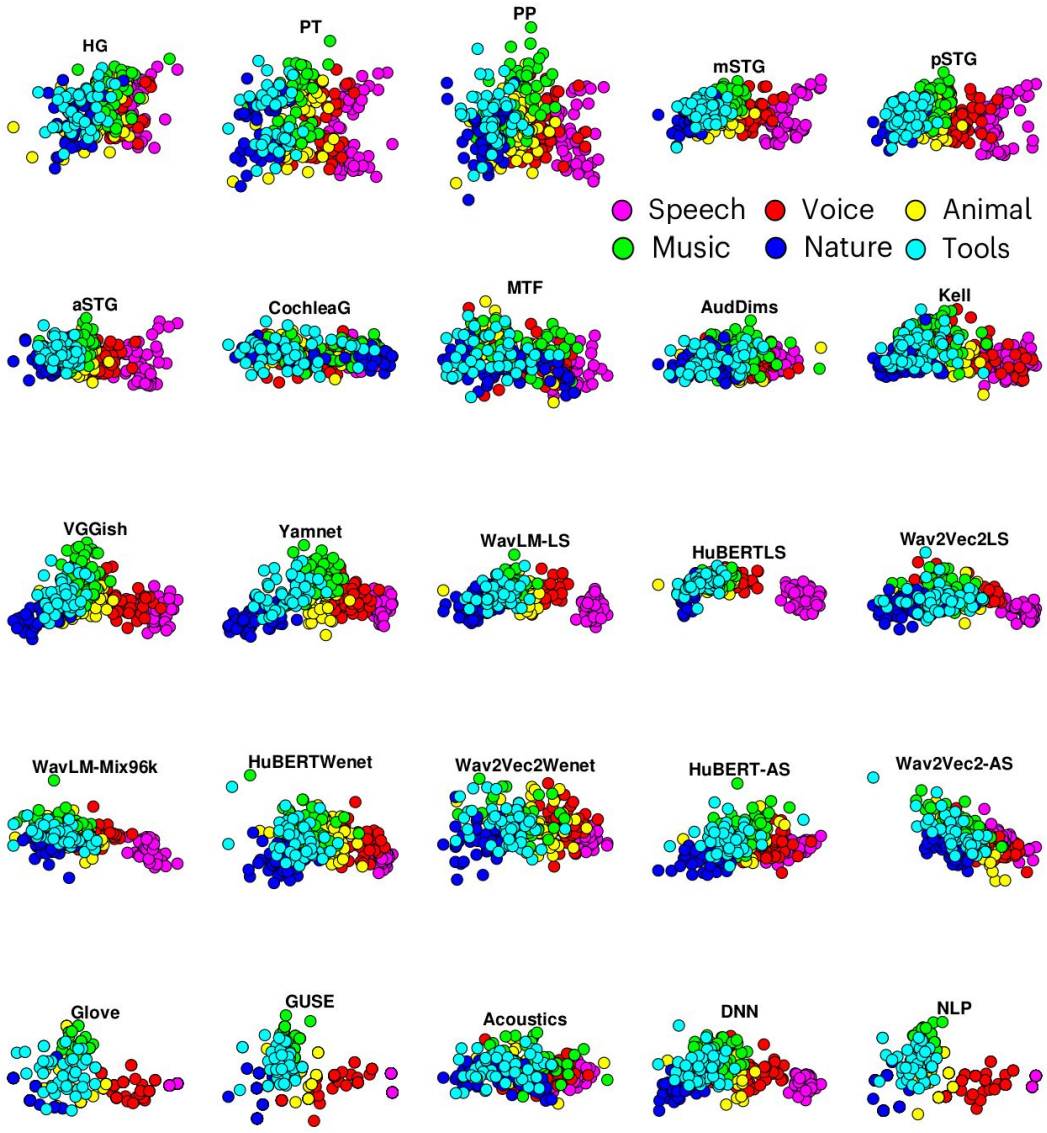

Figure 9: Metric multidimensional scaling (MDS) visualizations of natural-sound representations across the six auditory fMRI regions of interest (ROIs), acoustic models, semantic models, supervised DNNs, self-supervised DNNs, and combined model representations. The six ROIs comprise primary auditory cortex (HG: Heschl's gyrus), secondary auditory cortex (PT: planum temporale; PP: planum polare), and tertiary auditory cortex (mSTG, pSTG, and aSTG: middle, posterior, and anterior superior temporal gyrus, respectively). MDS was performed on standardized representational distances averaged across the analyzed components of each model. For visualization, MDS solutions were Procrustes-aligned using translation and rotation to the solution obtained from the average pSTG representational distances.

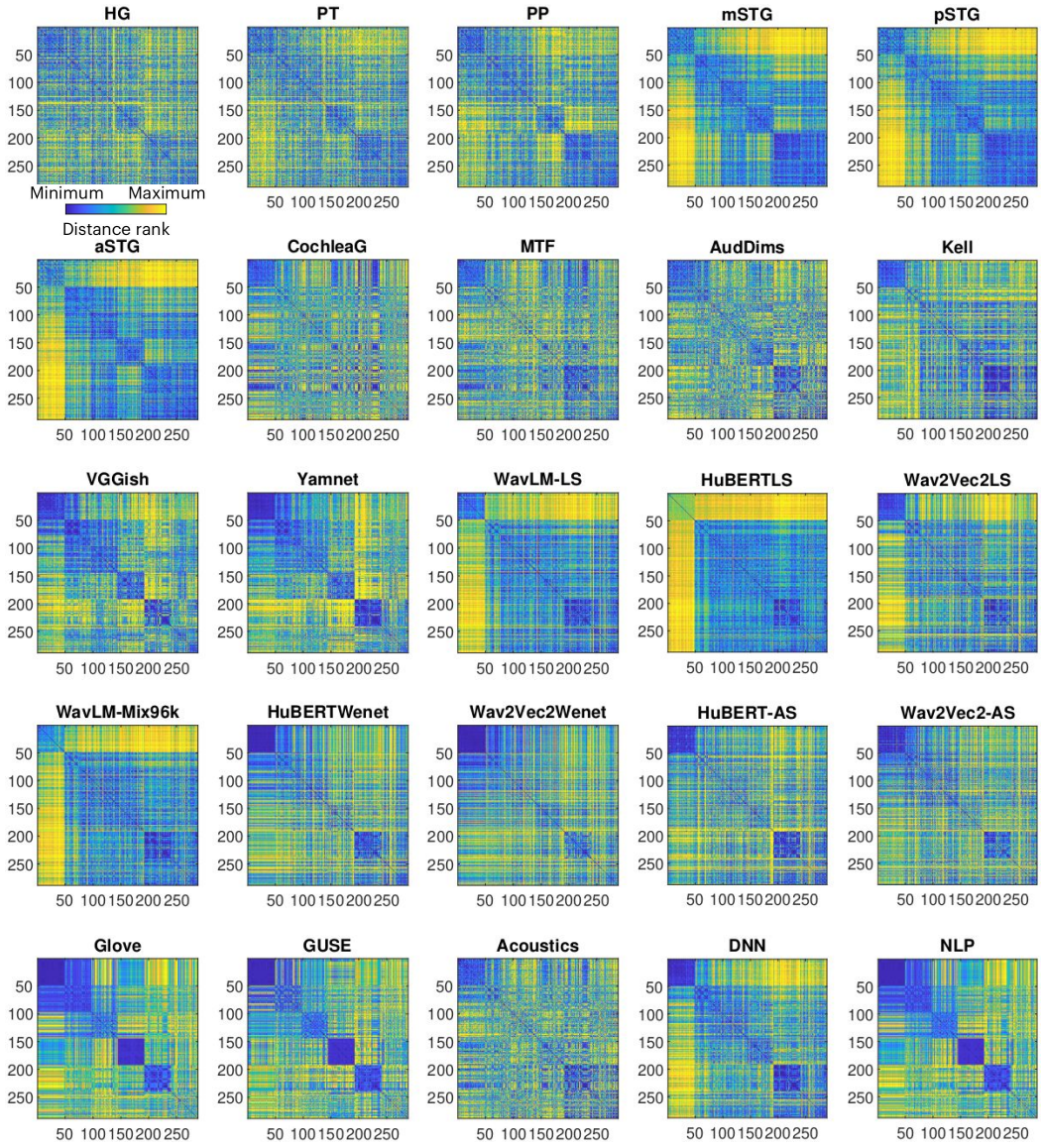

Figure 10: Representational dissimilarity matrices (RDMs) corresponding to the auditory fMRI ROIs and computational-model representations visualized with MDS in Fig. 9. RDMs are shown for the six auditory ROIs, acoustic models, semantic models, supervised DNNs, self-supervised DNNs, and combined model representations.

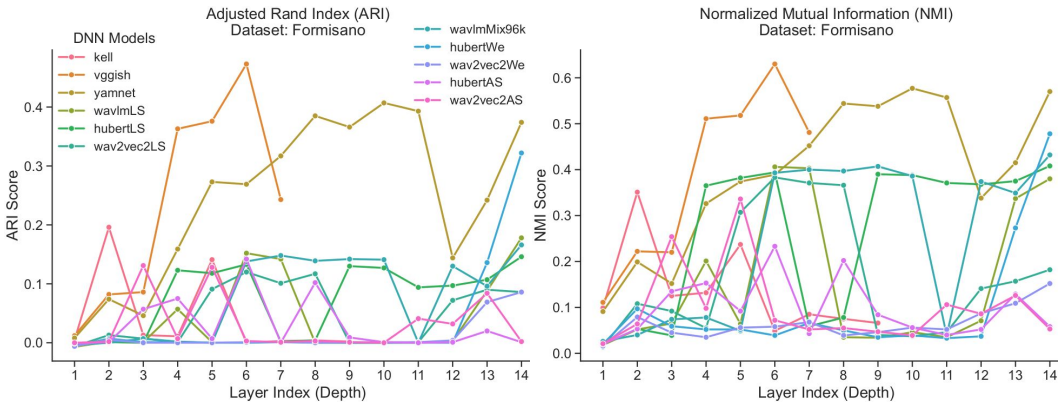

Figure 11: Category-structure correspondence across DNN layers. Adjusted Rand Index (ARI; left) and Normalized Mutual Information (NMI; right) quantify the correspondence between model-derived cluster assignments and the reference sound-category labels across layers of the supervised and self-supervised DNNs. Higher values indicate greater correspondence with the reference category structure.