Nature Communications

https://doi.org/10.1038/s41467-025-66731-7

Article in Press

Text-to-music generation models capture musical semantic representations in the human brain

Received: 27 May 2024

Accepted: 13 November 2025

Cite this article as: Denk, T.I., Takagi, Y., Matsuyama, T. *et al.* Text-to-music generation models capture musical semantic representations in the human brain. *Nat Commun* (2025). https://doi.org/10.1038/s41467-025-66731-7

Timo I. Denk, Yu Takagi, Takuya Matsuyama, Andrea Agostinelli, Tomoya Nakai, Christian Frank & Shinji Nishimoto

Text-to-music generation models capture musical semantic representations in the human brain

Timo I. Denk[1*†], Yu Takagi[2,3,4*†], Takuya Matsuyama[2,4], Andrea Agostinelli[1], Tomoya Nakai[5], Christian Frank[1], Shinji Nishimoto[2,4]

[1*]Google DeepMind, 1600 Amphitheatre Parkway Mountain View, CA 94043, USA.
[2]The University of Osaka, 1-1 Yamadaoka, Suita, Osaka, 565-0871, Japan.
[3]Nagoya Institute of Technology, Gokiso, Showa, Nagoya, Aichi, 466-8555, Japan.
[4]NICT, 4-2-1, NukuiKitamachi, Koganei, Tokyo, 184-8795, Japan.
[5]Araya Inc., 1-11 Sakuma, Kanda, Chiyoda, Tokyo 101-0025, Japan.

*Corresponding author(s). E-mail(s): simssosdenk@gmail.com; yutakagi322@gmail.com;
†These authors contributed equally to this work.

Abstract

Aligning brain activity with machine-learning models offers a way to understand how the brain represents the world, and such models relate to the brain. In this paper, we explore the functional correspondence between music-specific information and brain activity. Using fMRI, we reconstruct music either via retrieval or by conditioning the MusicLM generator on brain data. The generated music resembles the musical stimuli that human subjects experienced with respect to semantic properties such as genre, instrumentation, and mood. Using an encoding modeling analysis, we also demonstrate that semantic information derived from music, as well as information derived from purely textual descriptions of music stimuli, are represented in a largely overlapped region around the auditory cortex. Furthermore, we show that features from MusicLM predicted brain activity within the primary auditory cortex more accurately than the non-music models that do not specifically focus on music.

1 Introduction

Music has universal significance and acts as a medium for expression and communication in diverse cultures. The representation of music within our brains has been a topic of interest in neuroscience. Previous studies have examined human brain activity, as captured by functional magnetic resonance imaging (fMRI), while participants listened to music. These studies discovered musical feature representations in the brain, such as rhythms [1], timbres [2, 3], emotions [4], and musical genres [5, 6]. While these approaches shed light on the processing of music-related information in the brain, they have not explicitly modeled the complex, high-level semantic information in music. This gap persists because, although recent research has revealed how the brain encodes high-level semantic information in other modalities like vision [7, 8], a corresponding method for extracting high-level semantic features from music remains undeveloped.

With the advent of text-to-music models, conditional generation of high-fidelity music has become feasible [9–13]. Recently, auto-regressive and diffusion-based models have significantly advanced the quality of synthesis in both music and broader audio generation. AudioLM [14] suggests autoregressively modeling a hierarchical tokenization scheme composed of both semantic and acoustic discrete audio representations. MusicLM [10] integrates the AudioLM framework with the joint music+text embedding model MuLan [15], enabling the generation of high-fidelity music conditioned on detailed text descriptions. Other Transformer-based methods encompass [16], [12], and [13], with the last of these leveraging the EnCodec audio codec [17]. Furthermore, diffusion model-based strategies for music generation are used by Riffusion [9], with the most recent advances proposed in [11, 12, 18, 19]. This development bridges the gap between our linguistic understanding of music and the actual creation of musical compositions. Thus, questions arise: How do the text and music embeddings used by these music generation models correspond to representations in the human brain? Furthermore, if a correspondence exists, is it possible to generate music directly from brain activity?

In this paper, we explore the functional correspondence between music-specific, high-level semantic information and brain activity. We find that different components of our music generation model are predictive of activity in the human auditory cortex. The distinction between low- and high-level audio representations within the auditory cortex is less pronounced than that observed in the visual cortex for visual stimuli. We also investigate whether such information is as widely distributed as in other modalities such as vision, and compare the strength of the correspondence to other general auditory models. Our results suggest that within the auditory cortex there is overlap in the voxels that are predictable from (a) purely textual descriptions of music and (b) the music itself. Finally, we explore the feasibility of generating music from brain activity scans with MusicLM [10]. We generate music from fMRI scans by predicting high-level semantically structured music embeddings and using a deep neural network to generate music from those features. Our evaluation indicates that the reconstructed music semantically resembles the original music stimulus. Although we employ MusicLM and its constituent components, our methodology is fundamentally adaptable to any music generator, provided that the generator can accommodate the conditioning on a dense embedding. Through control analyses, we verify that the

MusicLM features capture information that goes beyond just the genre of the music stimulus.

2 Results

We present and discuss the results of two complementary analyses: decoding, which generates music from brain activity, and encoding, which predicts brain activity from music features. The decoding analysis demonstrates our ability to reconstruct music from fMRI signals that closely resemble the original stimuli to which human subjects were exposed, when the fMRI data were recorded. The encoding analysis reveals how different musical representations map onto patterns of brain activity during music perception. We pre-process the fMRI dataset from [20] in the same manner as [6]. In this work, we augment the original dataset [20] by introducing English text captions. Details of the datasets and preprocessing are provided in Section 4.1.

2.1 MuLan and MusicLM

MuLan [15] is a joint text+music embedding model consisting of two towers, one for text ($\text{MuLan}^{\text{text}}$) and one for music ($\text{MuLan}^{\text{music}}$). The text tower is a BERT [21] model pre-trained on a large text corpus. For the audio tower, we use the ResNet-50 variant [22]. MuLan's training objective is to minimize a contrastive loss between the 128-dimensional embeddings produced by each tower for an example pair of aligned music and text. For example, the embedding of a rock song's waveform is supposed to be close to the embedding of the text *rock music* and far from *calming violin solo*. In this paper, if we refer to a MuLan embedding, we mean by default the embedding of the music tower.

MusicLM [10] is a conditional music generation model. Conditioning signals include – but are not limited to – text, other music, and melody. Supplementary Figure 1 visualizes the components of MusicLM in the context of fMRI decoding. In our decoding pipeline MusicLM is conditioned on a MuLan embedding which we compute based on an fMRI response. In MusicLM, music is generated in two consecutive stages. The first stage learns to map a MuLan embedding to a sequence of w2v-BERT tokens. These tokens used in MusicLM are extracted from a w2v-BERT [23] model's activations in the 7th layer, by clustering them with k-means. MusicLM's second stage maps the w2v-BERT tokens from the first stage and the MuLan embedding to acoustic tokens. These stem from a SoundStream [24] model's residual vector quantizer. The resulting tokens are converted back into audio using a SoundStream decoder. As in AudioLM [14], the second stage is split into a coarse and fine modeling stage. All three stages are implemented as Transformer models.

2.2 Decoding: Music Generation from Brain Activity

With decoding, we refer to the attempt to reconstruct the original music stimuli (to which a test subject was exposed) based on their recorded brain activity. This process can be subdivided into (1) predicting a music embedding based on fMRI data and (2) retrieving or generating music based on that embedding. See Section 4.2 for details.

In our setup, going from fMRI to music requires the prediction of an intermediate music embedding. The concrete choice of the music embedding represents a bottleneck for subsequent music generation. Only if we can predict music embeddings close to music embeddings of the original stimulus heard by the subject will we be able to generate music that is similar to the original stimulus with MusicLM. The quality of the generation is further constrained by what the embedding can capture.

The results obtained when predicting different types of embedding are in Table 1. We find that $\text{MuLan}^{\text{music}}$ embeddings can be more accurately predicted from fMRI signals than $\text{MuLan}^{\text{text}}$, w2v-BERT-avg, or SoundStream-avg embeddings. Note that avg suffix denotes that multiple embeddings were averaged along the temporal dimension. When predicting $\text{MuLan}^{\text{music}}$ embeddings, we observed the highest correlation as well as the highest identification accuracy. Three reasons may contribute to this: (1) The $\text{MuLan}^{\text{text}}$ and $\text{MuLan}^{\text{music}}$ embeddings, may be more closely aligned to the high level of abstraction represented in the fMRI data than the other embedding types. (2) The heuristics we applied to align w2v-BERT and SoundStream embeddings to the fMRI data may not be optimal and may require further analysis. (3) $\text{MuLan}^{\text{music}}$ embeddings have access to more musical information than $\text{MuLan}^{\text{text}}$ embeddings because they are computed from the audio signal. Based on this finding, the remaining decoding results use $\text{MuLan}^{\text{music}}$ embeddings (for brevity referred to as MuLan embeddings) as a prediction target.

2.3 Quantitative Evaluation of Decoding Analysis

Figure 1 shows the main quantitative results of the music reconstruction. We use two quantitative measures to evaluate the similarity of generated music and original stimulus: identification accuracy (for two embeddings of different semantic levels of abstraction) and AudioSet top-n class agreement.

The quality limit imposed by the retrieval from the Free Music Archive (FMA) [25] is estimated using an oracle predictor. It simply bypasses the linear regression and instead retrieves an FMA clip based on the original music stimulus's MuLan embedding. It simulates the retrieval performance that we would achieve if our fMRI to MuLan prediction was perfect. The performance achieved by a model randomly sampling from FMA is indicated by the chance result in the plots.

Overall, we observed above-chance performance in all metrics, supporting our ability to extract musical information from the fMRI scans. The identification accuracy comparison across different embedding types hints at our generation to be most faithful to the original stimulus with respect to high-level semantic features, as captured by MuLan. Although this may seem unsurprising, given that MuLan is the target embedding of our prediction, it is not necessarily given that high-level semantic information about perceived music is contained in the brain response recorded in the first place. We also observed consistent prediction performance across different genres as labeled in the GTZAN dataset. The highest accuracy is achieved in classical music, which is likely due to its distinctive musical style. Additionally, for music genres containing vocals, we evaluate whether the gender of the vocalists is accurately reconstructed. The results indicate that gender-related properties are adequately captured in the reconstructions (detailed in Section 2.4). We also conduct human evaluations.

Four participants were asked to rate a total of 60 reconstructed songs along with randomly selected reconstructed music. The overall accuracy of the human evaluations is $77.33\% \pm 4.30$ in identifying the reconstruction. This accuracy closely matches the automatically calculated accuracy derived from extracted features, suggesting that our nonhuman identification analysis aligns with human perception.

Whether or not low-level information is contained in the generation is measured by the identification accuracy on w2v-BERT-avg. However, note that the results are confounded by the embedding averaging we perform for w2v-BERT to align the temporal resolution of the embeddings with that of the fMRI scans. Still, the numbers are in line with the qualitative observation we make, that is, low-level, acoustic features are relatively not well aligned, whereas the overall music style is. For instance, as described later, a beats per minute (bpm) analysis revealed only chance-level correspondence between originals and reconstructions.

The two ways to obtain music that we compare are retrieval from FMA and generation with MusicLM. Supplementary Figure 2 contains qualitative results for Subject 1, comparing the original stimulus with the music retrieved from FMA based on a predicted MuLan embedding and the music clips sampled from MusicLM. Note that sampling multiple clips and examining their differences is one way to qualitatively determine which information MusicLM adds to what the MuLan embedding contains. We find that both the retrieved and generated constructions are semantically similar to the original stimulus, e.g., with respect to genre, vocal style, and overall mood. The temporal structure of the stimulus is often not preserved in the generation. There are also failure cases in which the generation is from an entirely different genre. We observe that the generated music contains vocals, but the lyrics are mostly unintelligible and resemble mumbling. This is likely due to MusicLM underfitting the data, that is, not having enough modeling capacity to learn how to output coherent language.

2.4 Additional Decoding Results

To assess what information decoding can capture and to what extent, we perform a series of analyses.

In Supplementary Figure 3 we perform a comparison of retrieval and generation across the five different subjects. The main finding is that, qualitatively, the generation is overall consistent in quality across all five subjects. This is not necessarily a given when dealing with fMRI data, because of differences in subjective experiences, and suggests that our method is robust.

In Supplementary Figure 5 we calculate the confusion matrix for the decoding analysis. From this analysis, we observe that jazz tends to be confused with classic or blues, disco with hip-hop, and vice versa. We find a strong correlation between the pairwise distances in the PCA space (derived from the encoding analysis) and the identification accuracy in the confusion matrix (from decoding), which is intuitive (see the scatter plot), as shown in Supplementary Figure 6.

To evaluate the accuracy of decoding vocal gender, we perform a human annotation task. Specifically, we recruited one participant, who labeled all original and reconstructed songs as Male, Female, or Unknown. Jazz and Classical genres were excluded

because the original stimuli contained no vocal parts. The average classification accuracy across all subjects is 0.6667 ± 0.039 (mean $\pm$ s.t.d.; Subject 1 is 0.64583, Subject 2 is 0.70833, Subject 3 is 0.70833, Subject 4 is 0.625, and Subject 5 is 0.64583).

We estimate bpm for both original and reconstructed clips using the algorithm from Ellis [26] (as implemented in librosa's `beat_track` function). Computed across all test excerpts, the percentage of reconstructions with matching bpm was the same as that for a baseline sampling a random reconstructed track's bpm. A match is defined as the relative difference of the two bpm values being below ϵ, i.e.,

$$\text{bpm\_match} = \frac{\max(\text{bpm}_1, \text{bpm}_2)}{\min(\text{bpm}_1, \text{bpm}_2)} - 1 < \epsilon, \qquad \epsilon = 5\%. \tag{1}$$

Plausible causes of that are (1) noise in the bpm annotations, (2) lack of information in the temporally coarse fMRI signal, and (3) the MuLan embedding bottleneck. Future work could combine the MuLan embeddings with a complementary embedding that is known to encode the bpm well or attempt a bpm prediction directly from the fMRI signal.

2.5 Encoding: Predicting Brain Activity from Music

With encoding, we refer to modeling how the music generation model's internal representations map onto patterns of brain activity. Specifically, we predicted voxel-wise fMRI responses from embedded audio or text. Training data preparation is performed in the same manner as in decoding (outlined in Section 4.2). The modeling problem is inverse to that of decoding, that is, fMRI responses are predicted based on different embeddings. Model weights are estimated from training data using L2-regularized linear regression and subsequently applied to test data. We train our encoding model for the whole brain, not limited to specific region of interests. To examine whether nonlinearity improves performance, we additionally trained multilayer neural networks on the same data; however, they did not surpass the linear baseline (See Section 4.3 for details). We estimate the weights of the model from training data, and regularization parameters are explored during the training using a five-fold cross-validation. For evaluation, we use Pearson's correlation coefficients between predicted and measured fMRI signals. We compute statistical significance using blockwise permutation testing described in Section 4.1. For $\text{MuLan}^{\text{text}}$ and one-hot music genre vectors, we perform up-sampling to match $\text{MuLan}^{\text{music}}$'s sampling rate. See Section 4.3 for details.

Figure 2a shows the prediction accuracy of the encoding models for different types of audio-derived embeddings of music within MusicLM: MuLan and w2v-BERT-avg. It shows that both MuLan and w2v-BERT-avg predicted human brain activity in the auditory cortex. Given that the text-music model used in this study is not brain-inspired, it is intriguing that this correspondence with the brain emerge. In addition, although each embedding represents different levels of audio information from low (w2v-BERT-avg) to high (MuLan), they predicted fairly similar brain regions within the auditory cortex. Figure 2b further confirms that well-predicted voxels are largely overlapping between two embeddings. These results suggest that, unlike the visual cortex [8], hierarchical functional differentiation of audio-derived embeddings in the auditory cortex is not as pronounced as previously thought. Note that the results are

confounded by the embedding averaging we perform for w2v-BERT as described in Discussion. We provide additional results for all subjects in Supplementary Figure 10.

2.6 Comparison between Audio- and Text-derived MuLan Embeddings

We next investigate how much text-derived, abstract information about music is differently represented in the auditory cortex compared to audio-derived embeddings.

Figure 3a shows the prediction performance of the encoding models for $\text{MuLan}^{\text{text}}$ versus $\text{MuLan}^{\text{music}}$. We provide additional results for all subjects in Supplementary Figure 10. It shows that for some subjects, the inner side of the superior temporal sulcus represents music stronger than the outer side. However, there still seems to be modest functional differentiation in the brain. Although these two representations are trained to match [15], due to the many-to-many nature of text and music pairings, the objective cannot be achieved perfectly. We show that from a neuroscience perspective, $\text{MuLan}^{\text{music}}$ and $\text{MuLan}^{\text{text}}$ actually acquired fairly similar representations. Figure 3b further confirms that well-predicted voxels are largely overlapping between two embeddings. Although $\text{MuLan}^{\text{music}}$ and $\text{MuLan}^{\text{text}}$ predict similar voxels, they may represent different aspects of each voxel's activations. To explore this, we combine $\text{MuLan}^{\text{music}}$ and $\text{MuLan}^{\text{text}}$ into a single model. We then examine their distinct contributions by mapping the unique variance explained by each feature onto the cortex, utilizing banded ridge regression [27]. Supplementary Figure 8 shows that $\text{MuLan}^{\text{text}}$ exhibits less unique explained variance compared to $\text{MuLan}^{\text{music}}$. Again, this observation differs notably from trends in visual neuroscience [8], where text representations have distinct prediction performance. In summary, although $\text{MuLan}^{\text{text}}$ predicts many areas within the auditory cortex, most of the information contributing to this prediction is contained in $\text{MuLan}^{\text{music}}$.

2.7 Principal Component Analysis of the Encoding Model Weight Matrix

To further examine information representation in representative voxels, we perform principal component analysis (PCA) on the weight matrix of encoding model of $\text{MuLan}^{\text{music}}$ (Figure 3; see 4.3 for details and Supplementary Figure 10 for all subjects results). The top three PC components explained 25.8, 11.0 and 5.4% of the total variance, respectively. We observe a gradient from PC1 to PC2 or PC3 from the medial to lateral axis within auditory cortex, which is not apparent from comparisons between $\text{MuLan}^{\text{music}}$ and w2v-BERT. While we can observe this gradient consistently across subjects, there is also notable individual differences regarding the distribution of PC2 and PC3. To interpret the PCs, we also project each music onto a PC axes (Figure 3d). The music projected onto the three-dimensional PC space is rather intricately distributed, although there are some coarse clusters according to genres (e.g. Hip-hop and Classical are at opposite ends of the spectrum).

2.8 Comparison with Other Audio Models

In addition to MuLan$^{\text{music}}$ and w2v-BERT, we develop separate encoding models using three different approaches - wav2vec 2.0 [28], HuBERT [29], and melspectrogram. Both wav2vec 2.0 and HuBERT are transformer-based encoders trained through self-supervised learning on 960 hours of LibriSpeech. For our study, we utilize the hubert-large-ls960-ft and wav2vec2-large-960h models available on Huggingface. We employ the outputs from the last hidden layer of both architectures. The melspectrogram is computed using Librosa. All other settings remain consistent with those used in MuLan$^{\text{music}}$. A correlation analysis reveals very small linear correlations between MuLan$^{\text{music}}$ and melspectrogram features (see the Supplementary Figure 7). This suggests that each MuLan embedding encodes abstract musical semantics beyond low-level acoustics.

Figure 4 shows the prediction performance of different models across all five subjects. Overall, MuLan$^{\text{music}}$ predicts brain activity within the primary auditory cortex more accurately than the other audio models, including the large self-supervised audio models previously utilized in auditory neuroscience studies (prediction performance of MuLan$^{\text{music}}$ is significantly higher than that of HuBERT for all but Subject 2; Subject 1, $p = 0.002$; Subject 2, $p = 1.0$; Subject 3, $p = 0.006$; Subject 4, $p = 0.03$; Subject 5, $p = 0.00001$; two-sided paired t-test, family-wise error corrected).

2.9 Generalization Beyond Music Genre

We next investigate whether our model can generalize to the music genre that was not used during training. To do so, we ablate one genre during training and determine the identification accuracy on clips of the held-out genre (from the test set). Figure 5a shows that our model performs significantly above chance on the unseen music genres. Furthermore, we perform identification accuracy analyses restricted to within-genre comparisons. The classification accuracy is $0.716\% \pm 0.013$ (average $\pm$ s.e.m. across five subjects). While this is slightly lower than in the cross-genre condition as expected, the classification accuracy was well above 50% for all subjects. These results suggest that our music reconstruction method is capturing musical details well beyond the genre.

We further compare the prediction performances of the MuLan$^{\text{music}}$ and genre models to test whether our encoding model captures information beyond music genres. We find that, compared to the music genres vectors, the MuLan$^{\text{music}}$ embeddings predicted activity in the auditory cortex more broadly and with higher accuracy (Figure 5b). This is another piece of evidence that our model might predict beyond the music genre information. We provide additional results for all subjects in Supplementary Figure 11.

3 Discussion

One of the key goals in neuroscience is to understand the representations that govern the relationship between brain activity and sensory and cognitive experiences. To this end, researchers construct encoding models to quantitatively describe which features of these experiences (e.g., color, motion, and phonemes) are encoded as brain activity.

In contrast, they also build decoding models to infer the experienced content from a specific pattern of brain activity (for a review, see [30]).

In particular, in recent years, researchers have discovered correspondences between the internal representations of deep learning models and those of the brain in various sensory and cognitive modalities [31, 32]. These findings have advanced our understanding of brain functions through the development of (a) encoding models, which interpret representations based on their correspondence with brain functions [8, 33, 34] and (b) decoding models, which generate experienced content (such as visual images) from brain activity [8, 35–37]. More specifically, in the context of investigating auditory brain functions, researchers have developed encoding models using deep learning models that process auditory inputs [32], and conducted studies to generate perceived sounds from brain activity [38, 39].

Studies on music in the field of fMRI have focused on either decoding or encoding, often relying on hand-crafted features for the analysis [1, 2, 40, 41]. An exception is the work of Güçlü et al. [42], who built encoding models using CNN-based features optimized for music classification, revealing hierarchical representations in the auditory cortex. Our study extends these prior works by explicitly exploring semantic (text-based) representations and low-level (audio-based) representations of music in the brain using modern text-to-music generative models like MuLan and MusicLM. Unlike classification-focused models such as those used by Güçlü et al. [42], we leverage generative AI to investigate how human brain activity aligns with text- and music-based embeddings. Specifically, we quantitatively compare text- and music-based embeddings within the same model, demonstrating their correspondence to brain activity localized in the auditory cortex. Although higher-order auditory areas have been linked to melodic contour and note-surprisal coding [43], this localization might seem at odds with theories that emphasize distributed, modality-independent semantic hubs [41, 44]. However, we note that we detect a modest number of significant voxels in the lateral prefrontal, lateral temporal, and inferiorparietal cortex (see Supplementary Figure 9), showing that the representation is anchored in the auditory regions but not completely confined to them. We therefore interpret our results as indicating that high-level musical semantics are centered on, but not exclusive to, auditory cortices. We also show that semantic embeddings derived from music-specialized models outperform those from state-of-the-art speech models, highlighting the brain's specialization for music semantics. To do so, we introduce a method for music retrieval and generation by employing the embeddings of text-to-music generative model (MusicLM), and demonstrate that our music-specialized model outperforms other state-of-the-art speech models in modeling human brain activity while subjects are listening to music. Furthermore, unlike in the visual cortex, the hierarchical functional differentiation of audio-derived embeddings in the auditory cortex is less pronounced than previously thought.

Our study is positioned as a contribution to the field of NeuroAI [45], aiming to explore the similarities between representations in the human brain and AI systems. To achieve this, we adopt a dual strategy, employing both encoding and decoding methods in parallel as complementary tools to investigate the correspondence between the human brain and AI systems. This contrasts with many previous studies that focused

on either encoding or decoding alone. Encoding, while effective at modeling brain activity, does not intuitively reveal how much or what type of information is being represented. On the other hand, our decoding approach, though more interpretable, does not guarantee a precise mapping and struggles to provide a quantitative picture of how different features are differently represented to the brain. By combining both approaches, we provide a comprehensive perspective on how brain activity corresponds to musical features within modern generative AI. Although we include multiple control analyses to ensure robustness, interpreting by single encoding or decoding results is not enough, and our dual-strategy offers deeper insights into the relationship between the brain and AI.

The present study has several limitations, particularly those related to the modeling framework. First, the music generation model we use converts a MuLan embedding into music. Information that is not contained in the embedding, but required to produce a music clip, is added by the model. In the retrieval case, it is even guaranteed that the generation is musical, because it is directly pulled from a dataset of music. While this leads to human-digestible results, it also suggests a higher level of generation quality than there may be. Second, the main three factors limiting the generation quality are: (1) how much information we can extract with linear regression from the fMRI data, (2) what the chosen music embedding – in our case MuLan – can capture, and (3) the limitations of the music retrieval (dataset size and diversity) or generation (generative model capabilities). In the encoding analysis in Section 2.5 we investigate the limitations of (1 + 2) together by predicting brain activity for different embedding types. We disentangle (2 + 3) from (1) by showing the maximum attainable performance as an oracle dashed-line in Figure 1. It shows, for example, that top-1 class agreement on moods cannot exceed 78% given the MuLan embedding and FMA retrieval dataset choice. To observe the variability of (3), we experiment with three different generation processes, i.e., two sizes of FMA for retrieval and a generative model. Further investigation of (1), (2), and (3) remains an open challenge.

Beyond the modeling aspects discussed above, there are also limitations related to the neuroimaging data and human factors. The coarse temporal sampling rate of fMRI (1.5s, in the present study) is a limitation of the present study. However, it is noteworthy that even at the fMRI sampling rate of 2.5s, [38] showed temporal specificity at 200ms by using multi-voxel patterns. Different voxels might have information about different frequency bands, therefore, it is possible that broader temporal dynamics, such as changes in overall mood or instrumentation over a 15-second span, are represented in the brain and reflected in the reconstructions. MuLan embeddings also may encode these coarse-grained temporal features, enabling their reconstruction from the decoded embeddings. However, determining whether these apparent temporal similarities are coincidental or systematically captured by the model requires further analysis. This question is beyond the current scope of our study. How much information can actually be retrieved from fMRI is a subject for future research (see [46] for generating perceived natural movies from fMRI data using frequency-decomposed voxel-wise representations). Similarly, the temporal averaging of SoundStream and w2v-BERT embeddings likely worsens their expressiveness. Averaging these high-resolution features over the fMRI sampling window may have significantly reduced their ability

to capture fine-grained temporal dynamics. Instead of averaging, using nonlinear temporal alignment between model embeddings and brain activity could provide an alternative way to preserve temporal dynamics. Further exploration of such alignment techniques remains an important direction for future research. Fourth, the captions used in this study were written by individuals other than the fMRI participants, which may introduce a subjective mismatch between the subject's personal experience of the music and the descriptions provided. Although the use of MuLan text embedding achieves a reasonable level of encoding performance, this approach may abstract or omit participant-specific details of the musical experience.

There are several directions for future work. First, we perform music generation from fMRI signals that were recorded while human subjects were listening to music stimuli through headphones. A next step is to attempt the generation of music or musical attributes from a subject's imagination. In such a setting, a subject imagines a music clip they know well and the decoding analysis examines how faithfully the imagination can be generated. This has not been explored yet and would qualify for an actual mind-reading label. Second, comparing the generation properties among diverse subjects with different musical expertise is another possible direction of exploration. Comparing, for example, the generation quality between subjects who are professional conductors or musicians and subjects with barely any self-reported music experience could give insight into differences in subjective perception. Furthermore, investigating cross-subject generalization could provide a deeper understanding not only of individual variability but also of the commonalities shared across subjects. Such an analysis could help identify universal patterns in how the brain represents musical features, as well as the unique aspects of individual perception. Third, extending these models to incorporate individual listener profiles or adapting them to include contextual factors (e.g., mood, familiarity with music) could help bridge this gap and deepen our understanding of the interplay between music, brain activity, and generative models. Lastly, while this study employs a transformer-based text-to-music model, MusicLM, this may not be the optimal architecture for modeling the human brain. In our analyses, we compare MusicLM with various models including low-level mel-spectrogram features and state-of-the-art speech models. However, future research should also explore other task-optimized DNN models [42], trained to solve human-relevant tasks, as used in prior work in this field. Finally, while we relied on MuLan embeddings for both decoding and encoding, our approach does not yet explore the potential benefits of combining high-level semantic features with low-level audio features for music reconstruction. Inspired by advances in vision-based generative models that integrate both types of information, future work could condition the generation process on both high-level and low-level features. Such an approach could better capture nuanced auditory details and improve reconstruction quality.

In this work we explored the research direction of generating music from recorded human brain activity. By conditioning MusicLM on a dense music embedding predicted from fMRI data, we were able to generate music which resembles the original music stimuli on a semantic level. We also investigated the connection between a text-to-music model and the human brain in a quantitative manner by constructing an

encoding model. Specifically, we assessed where and to what extent high-level semantic information and low-level acoustic features of music are represented in the human brain. Although text-to-music models are rapidly developing, their internal processes are still poorly understood. This study provides a quantitative interpretation from a biological perspective. Given the nascent stage of music generation models, this work is an initial step. Future work may improve the temporal alignment between generation and stimulus or explore the generation of music from pure imagination.

4 Methods

This section describes the datasets, preprocessing steps, and both decoding and encoding analyses conducted in this study.

4.1 Music fMRI Dataset

The dataset contains music stimuli from 10 genres (blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock) which were sampled randomly from the (music-only) dataset GTZAN [47]. A total of 54 music pieces (30s, 22.050kHz) were selected from each genre, providing 540 music pieces. A 15s music clip was selected at random from each music piece. The dataset contains 480 examples for training and 60 for reporting the final results.

Data collection details.

During scanning, five participants (referred to as Subject 1 to Subject 5; aged between 23 and 33 years old; 3 males) were asked to focus on a fixation cross at the center of the screen and to listen to the music clips through MRI-compatible insert earphones (Model S14, Sensimetrics). Every subject heard the same music clips. This headphone model can attenuate scanner noise and has been widely used in previous MRI studies with auditory stimuli [48]. Scanning was performed using a 3.0T MRI scanner (TIM Trio; Siemens, Erlangen, Germany) equipped with a 32-channel head coil. For functional scanning, we scanned 68 interleaved axial slices with a thickness of 2.0mm without a gap using a T2*-weighted gradient echo multi-band echo-planar imaging (MB-EPI) sequence [49] (repetition time (TR, aka. sampling interval) = 1,500ms, echo time (TE) = 30ms, flip angle (FA) = 62°, field of view (FOV) = $192 \times 192\text{mm}^2$, voxel size = $2 \times 2 \times 2\text{mm}^3$, multi-band factor = 4). A total of 410 volumes were obtained for each run. Informed consent was obtained from all participants prior to their participation. This experiment was approved by the ethics and safety committee of the National Institute of Information and Communications Technology in Osaka, Japan No statistical method was used to predetermine sample size. No data were excluded from the analyses. All participants engaged in the same set of experiments. The Investigators were not blinded to allocation during experiments and outcome assessment.

In addition to the fMRI experiment described above, four additional participants (1 female, 3 males; aged 21–25 years) were recruited for a human-rating experiment. Participants were asked to rate a total of 60 reconstructed songs together with randomly selected reconstructed music. Informed consent was obtained from all participants prior to their participation. All procedures were approved by the ethics and safety

committee of the National Institute of Information and Communications Technology in Osaka, Japan, and by the ethics committee of The University of Osaka, Japan.

Preprocessing details.

We preprocess the music genre neuroimaging dataset from [20] in the same fashion as [6], recited below. Motion correction is performed for each run using the Statistical Parametric Mapping toolbox (SPM8; Wellcome Trust Centre for Neuroimaging, London, UK; http://www.fil.ion.ucl.ac.uk/spm/). All volumes are aligned to the first EPI image for each participant. For each clip, 2s of fade-in and fade-out were applied and the overall intensity was normalized. Low-frequency drift is removed using a median filter with a 240s window. To augment model fitting accuracy, the response for each voxel is normalized by subtracting the mean response and then scaling it to the unit variance. We use FreeSurfer [50] to identify the cortical surfaces from the anatomical data and register them with the voxels of the functional data. We use only cortical voxels as targets of the analysis for each participant. For each participant, we use the voxels identified in the cerebral cortex in the analysis (53,421 to 64,700 voxels per participant).

Text captions.

In this work, we augment the original dataset [20] by introducing English text captions, which we have made publicly available. These captions, which average approximately 46 words or 280 characters in length, typically describe musical pieces in terms of genre, instrumentation, rhythm, and mood. They often comprise fragmented or semi-complete sentences, with an average of about 4.5 sentences per caption. The style of writing is subjective, reflecting not only the technical components of the music, such as the instruments used, but also the emotional responses or atmospheres they might evoke in listeners. The captions were written in Japanese and translated using DeepL. Several exemplar captions and the instructions given to the raters are in Section 1 of Supplementary Information. In this paper, captions are used to study how purely semantic text-derived embeddings relate to brain activity induced by the corresponding music stimuli.

4.2 Decoding: Generating Music from fMRI

With decoding, we refer to the attempt to reconstruct the original music stimuli (to which a test subject was exposed) based on their recorded brain activity. This process can be subdivided into (1) predicting a music embedding based on fMRI data and (2) retrieving or generating music based on that embedding.

Let $\mathbf{R} \in \mathbb{R}^{n \times s \times d_{\text{fmri}}}$ denote the response tensor (obtained by fMRI for each of the five participants), where $n = 540$ is the number of stimuli (that is, 15-second music clips), $s = 10$ is the number of fMRI scans per stimulus (15s) and d_{fmri} is the number of voxels. d_{fmri} varies slightly between subjects depending on the physical size of their brain. For Subject 1 it is around 60k. Our prediction targets are the music embeddings of the stimulus (e.g., MuLan), $\mathbf{T} \in \mathbb{R}^{n \times r \times d_{\text{emb}}}$, with r being the number of embeddings computed per 15s clip (which depends on the embedding model's window size and the constant step size of 1.5s by which we advance this window). Supplementary Table 1

lists the embeddings, derived from models present in MusicLM, that we consider as candidate music embeddings in our architecture.

To align $\mathbf{R}$ and $\mathbf{T}$ along the time dimension, we average entries in $\mathbf{R}$ in a sliding-window fashion to match the time ranges for which feature vectors in $\mathbf{T}$ were computed. For example, to predict the MuLan embedding ranging from 0s to 10s (due to MuLan's window size being 10s) we rely on the average of five fMRI scans (from 0-1.5s, 1.5-3s, ..., 9-10.5s). This leaves us with $m := n \times r$ pairs of response and target, which we split following [20]. We model the relationship between music embeddings and responses with weight matrix $\mathbf{W} \in \mathbb{R}^{d_{\text{fmri}} \times d_{\text{emb}}}$:

$$\hat{\mathbf{T}} = \mathbf{R}\mathbf{W}\,. \tag{2}$$

We use an L2-regularized linear regression to estimate $\mathbf{W}$ on the training dataset. Note that there is no generalization between different subjects, because of anatomical differences. For each subject and target dimension, the regression regularization hyperparameter is tuned with five-fold cross validation on the training dataset, while a test split is held out for later evaluation. We detail this process below.

Hyperparameter Tuning

We use a ridge regression to estimate the parameters of our model [27, 51]: Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be a feature matrix with n samples and p features, $\mathbf{y} \in \mathbb{R}^n$ a target vector, and $\alpha > 0$ a fixed regularization hyperparameter. Ridge regression defines the weight vector $\mathbf{b}^* \in \mathbb{R}^p$ as:

$$\mathbf{b}^* = \text{argmin}_{\mathbf{b}} \|\mathbf{X}\mathbf{b} - \mathbf{y}\|_2^2 + \alpha \|\mathbf{b}\|_2^2\,. \tag{3}$$

The equation has a closed-form solution $\mathbf{b}^* = \mathbf{M}\mathbf{y}$, where $\mathbf{M} = (\mathbf{X}^\top \mathbf{X} + \alpha \mathbf{I}_p)^{-1} \mathbf{X}^\top \in \mathbb{R}^{p \times n}$.

To determine α, we run a 5-fold cross-validation on the training data. Note that there is one parameter α per regression target, that is, 128 in our case when predicting MuLan embeddings. We inspect the performance on training and evaluation data around the chosen vector α (opt) in the Supplementary Figure 13.

Training is performed independently for each anatomically defined region of interest (ROI) of the Destrieux atlas [52]. An ROI is a group of voxels. From all 150 ROIs we select the top $n_{\text{ROIs}} = 6$ with the highest correlation scores (as determined by cross-validation) and create an ensemble model by averaging their predictions. Concretely, this leaves us with ROIs that vary in size. ROI sizes vary between subjects. The top six ROIs of Subject 1, for example, which are the most predictive of MuLan embeddings, have dimensionalities 61, 70, 109, 218, 296, and 706. The median across all subjects is 138.5 voxels; the average is 258.6 voxels. Although the exact location of the ROIs chosen for each subject varies, for all subjects, the ROIs are chosen primarily from the auditory cortex (See Table 2 for a complete list of the selected ROIs.). For a given 15s music stimulus, we predict r many embeddings (depending on the chosen embedding type).

Selected ROIs for the Decoding

We explore two different approaches to derive the original stimulus from the prediction $\hat{\mathbf{T}}$ namely retrieving similar music from an existing music corpus and generating music with MusicLM.

For the retrieval we compute the MuLan embeddings for the first 15s of every music clip in the Free Music Archive (FMA) [25]. Unless stated otherwise, we use the large variant which contains a wide range of diverse music; concretely, 106,574 music tracks from 161 unbalanced genres. The retrieved music is the audio of the clip whose embeddings are the closest to the predicted one as measured by the cosine similarity. While it is difficult to completely verify that none of the stimuli in the FMA overlap with those in the GTZAN dataset, we note that the FMA dataset is not used in the training of our models.

Alternatively, we can generate the music by conditioning the MusicLM model (a high-level overview is in Section 2.1) on the predicted embeddings. The model can then be used to generate music conditionally. To condition MusicLM we average the $r = 4$ predicted MuLan embeddings along the time dimension. This is not strictly necessary, but we empirically found the generated outputs to be more stable compared to a version in which we provided all four embeddings to the model.

The two methods, retrieval and generation, have different advantages and disadvantages. The retrieval approach constrains the faithfulness by its limited size. The predicted embedding could potentially contain rich information about the song which is partially lost by mapping it onto its nearest neighbor in the dataset. The generative model, on the other hand, can in theory generate any kind of music covered by its training distribution, making it conceptually more powerful. That includes tracks which are not training examples (e.g., combinations of musical concepts). A disadvantage of this method is that the generation model may not adhere precisely to the provided embedding.

Following the decoding literature [8, 39], we compute an identification accuracy of the predicted d-dimensional embeddings with respect to their target embeddings.

Assume that there is a matrix of predicted embeddings $\mathbf{P} \in \mathbb{R}^{n \times d}$ and a matrix (of equal size) that contains target embeddings $\mathbf{T}$. Let $\mathbf{C} \in \mathbb{R}^{n \times n}$ be computed from $\mathbf{P}$ and $\mathbf{T}$, specifically $C_{i,j}$ is the Pearson correlation coefficient between the i-th row of $\mathbf{P}$ and the j-th row of $\mathbf{T}$.

The identification accuracy for the i-th prediction is defined as:

$$\text{id acc}_i = \frac{1}{n-1} \sum_{j=1}^{n} \mathbb{1}[C_{i,i} > C_{i,j}] \ . \tag{4}$$

The identification accuracy for all examples is simply the average:

$$\text{id acc} = \frac{1}{n} \sum_{i=1}^{n} \text{id acc}_i \ . \tag{5}$$

This score, ranging from 0 to 1 with 0.5 indicating performance equivalent to random chance, provides a quantified measure of how well an embedding was predicted in relation to other embeddings in the dataset.

An intuitive view on the identification accuracy is that a model with 86% such accuracy, on average, 14% of the candidates retrieved will score higher, i.e., have a higher correlation coefficient than the correct candidate. In a dataset with 60 examples, the average rank of the correct music clip would be $0.14 \times 60 \approx 8$.

Top-n Class Agreement Metric

The algorithm used to calculate the top-n class agreement is presented below. It is the intersection-over-union among the top-n classes (of a certain group; see Section 2 of Supplementary Information) predicted by the LEAF [53] classifier operating on two music clips.

Algorithm 1 Calculate top-n class agreement

Input: $classesA$: ordered list of class labels for audio clip A
Input: $classesB$: ordered list of class labels for audio clip B
Input: n: number of top-n classes to compare
$topA \leftarrow$ first n elements of $classesA$
$topB \leftarrow$ first n elements of $classesB$
$intersection \leftarrow$ set of common elements in $topA$ and $topB$
$union \leftarrow$ set of all unique elements in $topA$ and $topB$
return $\frac{|\text{intersection}|}{|\text{union}|}$

As a second metric we also use top-n class agreement based on the LEAF [53] classifier operating on AudioSet classes [54]. In this context, we compute the per-class probabilities for original and generated music. We then look at three groups of music-related classes, namely genres, instruments, and moods. For each group, we compute the top-n agreement measuring how much overlap there is between the top-n most probable class labels computed for original and generated music. The full list of AudioSet class names in each group is in Section 2 of Supplementary Information.

4.3 Encoding: Whole-brain Voxel-wise Modeling

To interpret the internal representations of MusicLM, we examine the correspondence between them and recorded brain activity. We first build whole-brain voxel-wise encoding models to predict fMRI signals using different music embeddings occurring in MusicLM: audio-derived embeddings (MuLan$^{\text{music}}$ and w2v-BERT-avg), and text-derived embeddings (MuLan$^{\text{text}}$). The text-derived embeddings are particularly interesting to study, because they can – by definition – only represent the high-level information contained in the music caption they are computed from. The MuLan$^{\text{text}}$ embeddings we use have a 1:1 correspondence to GTZAN clips and are computed by inferring MuLan's text tower (a fine-tuned BERT model) for a given GTZAN clip's text caption. Caption examples are provided in Section 1 of Supplementary

Information. Furthermore, to investigate the unique contribution of MuLan$^{\text{music}}$ and MuLan$^{\text{text}}$, we use variance partitioning analysis. Unique contribution is calculated as the difference between the total variance explained (defined by Pearson's correlation coefficients) by the full model and another model that removed a feature (i.e. MuLan$^{\text{text}}$ or MuLan$^{\text{music}}$).

We also compare the prediction performance of MuLan$^{\text{music}}$ with other audio-derived features for anatomically defined primary auditory cortex (A1) [55]. The melspectrogram was used as a low-level feature, while HuBERT [29] and wav2vec2.0 [28] were used as high-level features. A key difference between MuLan$^{\text{music}}$ and HuBERT/wav2vec2.0 is that MuLan$^{\text{music}}$ is trained on music, whereas the latter are trained on speech. We detail each embedding below. In addition, to determine whether MuLan$^{\text{music}}$ encompasses more than genre information, we compare the prediction performance of the MuLan$^{\text{music}}$ model versus one-hot vectors of music genre for each GTZAN clip.

Principal Component Analysis

To further examine information representation in each voxel, we perform PCA on the weight matrix of encoding model of MuLan$^{\text{music}}$. For each of the top 600 voxels of predictive power (approximately 1% of all voxels, almost all in the auditory cortex), the weight matrix (voxels $\times$ 128) estimated by the MuLan encoding model was extracted. The weight matrices (voxels $\times$ 5 $\times$ 128) of the five participants were concatenated to obtain results for the whole group of participants. Dimensionality reduction of the weight matrices was performed using PCA on the feature direction of the concatenated weight matrices.

Voxel performance significance test

We shuffled the voxel's actual response time course, before taking the linear correlation between the predicted response time course and the permuted response time course. Repeating this process for 10,000 trials provided a null distribution for each voxel. We identified voxels with scores that were significantly higher than this null distribution. We set the threshold for statistical significance at $P < 0.05$ and corrected for multiple comparisons using the FDR procedure.

Non-linear Encoding Models

To test whether additional non-linearity benefits prediction, we train two compact non-linear regressors on primary auditory cortex voxels (the same protocol as Figure 4). The first is a two-layer multilayer perceptron (MLP) in which the MuLan vector passes through fully connected layers of 512 units with ReLU activations before a linear read-out to the voxel dimension. The second modeling approach projects the embedding to a 256-dimensional space and feeds it to a 2-layer, 4-head Transformer encoder (feed-forward width = 512, dropout = 0.1) followed by a linear read-out. Both networks are optimized with AdamW (Loshchilov and Hutter [56]; batch size 64, initial learning rate 3×10^{-4} with cosine decay to 1.5×10^{-4}, weight decay 0.02) for up to 50 epochs and stopped early after five epochs without validation improvement. As summarized in Supplementary Figure 12, neither model outperforms ridge regression

($p > 0.05$ between ridge vs. MLP [Subjects 1–5: $p = \{0.81, 1.0, 1.0, 1.0, 1.0\}$] and ridge vs. transformer [Subjects 1–5: $p = \{1.0, 0.43, 0.29, 1.0, 0.06\}$] after multiple corrections for all subjects). These findings suggest that the variance explainable from MuLan embeddings is largely linear in this region.

Data Availability

Data supporting the findings of this study are available at OpenNeuro.org (raw MRI data; (https://doi.org/10.18112/openneuro.ds003720.v1.0.0) and Kaggle (https://doi.org/10.34740/kaggle/ds/3535804). Audio examples can be found at https://google-research.github.io/seanet/brain2music.

Code Availability

The analysis code is provided at https://github.com/b2m-musiclm/b2m (https://doi.org/10.5281/zenodo.17487214). MuLan and MusicLM are proprietary software and not publicly available.

References

[1] Alluri, V., Toiviainen, P., Jääskeläinen, I.P., Glerean, E., Sams, M., Brattico, E.: Large-scale brain networks emerge from dynamic processing of musical timbre, key and rhythm. NeuroImage **59**(4), 3677–3689 (2012) https://doi.org/10.1016/j.neuroimage.2011.11.019

[2] Toiviainen, P., Alluri, V., Brattico, E., Wallentin, M., Vuust, P.: Capturing the musical brain with lasso: Dynamic decoding of musical features from fmri data. NeuroImage **88**, 170–180 (2014) https://doi.org/10.1016/j.neuroimage.2013.11.017

[3] Allen, E.J., Moerel, M., Lage-Castellanos, A., De Martino, F., Formisano, E., Oxenham, A.J.: Encoding of natural timbre dimensions in human auditory cortex. NeuroImage **166**, 60–70 (2018) https://doi.org/10.1016/j.neuroimage.2017.10.050

[4] Koelsch, S., Fritz, T., Cramon, D.Y., Müller, K., Friederici, A.D.: Investigating emotion with music: An fmri study. Human Brain Mapping **27**(3), 239–250 (2006) https://doi.org/10.1002/hbm.20180 https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbm.20180

[5] Casey, M.A.: Music of the 7ts: Predicting and decoding multivoxel fmri responses with acoustic, schematic, and categorical music features. Frontiers in psychology **8**, 1179 (2017)

[6] Nakai, T., Koide-Majima, N., Nishimoto, S.: Correspondence of categorical and feature-based representations of music in the human brain. Brain and Behavior **11**(1), 01936 (2021) https://doi.org/10.1002/brb3.1936 https://onlinelibrary.wiley.com/doi/pdf/10.1002/brb3.1936

[7] Wang, A.Y., Kay, K., Naselaris, T., Tarr, M.J., Wehbe, L.: Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. Nature Machine Intelligence **5**(12), 1415–1426 (2023)

[8] Takagi, Y., Nishimoto, S.: High-resolution image reconstruction with latent diffusion models from human brain activity. bioRxiv (2022) https://doi.org/10.1101/2022.11.18.517004 https://www.biorxiv.org/content/early/2022/11/21/2022.11.18.517004.full.pdf

[9] Forsgren, S., Martiros, H.: Riffusion - Stable diffusion for real-time music generation (2022). https://riffusion.com/about

[10] Agostinelli, A., Denk, T.I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Huang, Q., Jansen, A., Roberts, A., Tagliasacchi, M., et al.: Musiclm: Generating music from text. arXiv preprint arXiv:2301.11325 (2023)

[11] Huang, Q., Park, D.S., Wang, T., Denk, T.I., Ly, A., Chen, N., Zhang, Z., Zhang, Z., Yu, J., Frank, C., et al.: Noise2music: Text-conditioned music generation with diffusion models. arXiv preprint arXiv:2302.03917 (2023)

[12] Lam, M.W., Tian, Q., Li, T., Yin, Z., Feng, S., Tu, M., Ji, Y., Xia, R., Ma, M., Song, X., *et al.*: Efficient neural music generation. Advances in Neural Information Processing Systems **36**, 17450–17463 (2023)

[13] Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y., Défossez, A.: Simple and controllable music generation. Advances in Neural Information Processing Systems **36**, 47704–47720 (2023)

[14] Borsos, Z., Marinier, R., Vincent, D., Kharitonov, E., Pietquin, O., Sharifi, M., Roblek, D., Teboul, O., Grangier, D., Tagliasacchi, M., *et al.*: Audiolm: a language modeling approach to audio generation. IEEE/ACM transactions on audio, speech, and language processing **31**, 2523–2533 (2023)

[15] Huang, Q., Jansen, A., Lee, J., Ganti, R., Li, J.Y., Ellis, D.P.: Mulan: A joint embedding of music audio and natural language. arXiv preprint arXiv:2208.12415 (2022)

[16] Donahue, C., Caillon, A., Roberts, A., Manilow, E., Esling, P., Agostinelli, A., Verzetti, M., Simon, I., Pietquin, O., Zeghidour, N., et al.: Singsong: Generating musical accompaniments from singing. arXiv preprint arXiv:2301.12662 (2023)

[17] Défossez, A., Copet, J., Synnaeve, G., Adi, Y.: High Fidelity Neural Audio Compression (2022)

[18] Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Wang, W., Plumbley, M.D.: Audioldm: Text-to-audio generation with latent diffusion models. arXiv preprint arXiv:2301.12503 (2023)

[19] Ghosal, D., Majumder, N., Mehrish, A., Poria, S.: Text-to-audio generation using instruction-tuned llm and latent diffusion model. arXiv preprint arXiv:2304.13731 (2023)

[20] Nakai, T., Koide-Majima, N., Nishimoto, S.: Music genre neuroimaging dataset. Data in Brief **40**, 107675 (2022) https://doi.org/10.1016/j.dib.2021.107675

[21] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (long and Short Papers), pp. 4171–4186 (2019)

[22] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern

Recognition, pp. 770–778 (2016)

[23] Chung, Y.-A., Zhang, Y., Han, W., Chiu, C.-C., Qin, J., Pang, R., Wu, Y.: W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In: 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 244–250 (2021)

[24] Zeghidour, N., Luebs, A., Omran, A., Skoglund, J., Tagliasacchi, M.: Soundstream: An end-to-end neural audio codec. IEEE/ACM Transactions on Audio, Speech, and Language Processing **30**, 495–507 (2021)

[25] Defferrard, M., Benzi, K., Vandergheynst, P., Bresson, X.: FMA: A dataset for music analysis. In: 18th International Society for Music Information Retrieval Conference (ISMIR) (2017). https://arxiv.org/abs/1612.01840

[26] Ellis, D.P.W.: Beat tracking by dynamic programming. Journal of New Music Research **36**(1), 51–60 (2007) https://doi.org/10.1080/09298210701653344 https://doi.org/10.1080/09298210701653344

[27] Tour, T.D., Eickenberg, M., Nunez-Elizalde, A.O., Gallant, J.L.: Feature-space selection with banded ridge regression. NeuroImage **264**, 119728 (2022)

[28] Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems **33**, 12449–12460 (2020)

[29] Hsu, W.-N., Bolte, B., Tsai, Y.-H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A.: 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)

[30] Naselaris, T., Kay, K., Nishimoto, S., Gallant, J.: Encoding and decoding in fmri. NeuroImage **56**, 400–10 (2011) https://doi.org/10.1016/j.neuroimage.2010.07.073

[31] Yamins, D.L., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., DiCarlo, J.J.: Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proceedings of the national academy of sciences **111**(23), 8619–8624 (2014)

[32] Kell, A.J.E., Yamins, D.L.K., Shook, E.N., Norman-Haignere, S.V., McDermott, J.H.: A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. Neuron **98**(3), 630–64416 (2018) https://doi.org/10.1016/j.neuron.2018.03.044

[33] Güçlü, U., Gerven, M.A.: Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. Journal of Neuroscience

35(27), 10005–10014 (2015)

[34] Cross, L., Cockburn, J., Yue, Y., O'Doherty, J.P.: Using deep reinforcement learning to reveal how the brain encodes abstract state-space representations in high-dimensional environments. Neuron **109**(4), 724–7387 (2021) https://doi.org/10.1016/j.neuron.2020.11.021

[35] Shen, G., Horikawa, T., Majima, K., Kamitani, Y.: Deep image reconstruction from human brain activity. PLOS Computational Biology **15**(1), 1006633 (2019) https://doi.org/10.1371/journal.pcbi.1006633

[36] Chen, Z., Qing, J., Xiang, T., Yue, W.L., Zhou, J.H.: Seeing beyond the brain: Masked modeling conditioned diffusion model for human vision decoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)

[37] Takagi, Y., Nishimoto, S.: Improving visual image reconstruction from human brain activity using latent diffusion models via multiple decoded inputs. arXiv (2023). https://doi.org/10.48550/ARXIV.2306.11536 . https://arxiv.org/abs/2306.11536

[38] Santoro, R., Moerel, M., Martino, F.D., Valente, G., Ugurbil, K., Yacoub, E., Formisano, E.: Reconstructing the spectrotemporal modulations of real-life sounds from fMRI response patterns. Proceedings of the National Academy of Sciences **114**(18), 4799–4804 (2017) https://doi.org/10.1073/pnas.1617622114

[39] Park, J.-Y., Tsukamoto, M., Tanaka, M., Kamitani, Y.: Natural sounds can be reconstructed from human neuroimaging data using deep neural network representation. PLoS biology **23**(7), 3003293 (2025)

[40] Hoefle, S., Engel, A., Basilio, R., Alluri, V., Toiviainen, P., Cagy, M., Moll, J.: Identifying musical pieces from fmri data using encoding and decoding models. Scientific reports **8**(1), 2266 (2018)

[41] Alluri, V., Toiviainen, P., Lund, T.E., Wallentin, M., Vuust, P., Nandi, A.K., Ristaniemi, T., Brattico, E.: From vivaldi to beatles and back: predicting lateralized brain responses to music. Neuroimage **83**, 627–636 (2013)

[42] Güçlü, U., Thielen, J., Hanke, M., Van Gerven, M.: Brains on beats. Advances in Neural Information Processing Systems **29** (2016)

[43] Sankaran, N., Leonard, M.K., Theunissen, F., Chang, E.F.: Encoding of melody in the human auditory cortex. Science Advances **10**(7), 0010 (2024)

[44] Koelsch, S., Vuust, P., Friston, K.: Predictive processes and the peculiar case of music. Trends in cognitive sciences **23**(1), 63–77 (2019)

[45] Zador, A., Escola, S., Richards, B., Ölveczky, B., Bengio, Y., Boahen, K., Botvinick, M., Chklovskii, D., Churchland, A., Clopath, C., *et al.*: Catalyzing next-generation artificial intelligence through neuroai. Nature communications **14**(1), 1597 (2023)

[46] Nishimoto, S., Vu, A.T., Naselaris, T., Benjamini, Y., Yu, B., Gallant, J.L.: Reconstructing visual experiences from brain activity evoked by natural movies. Current Biology **21**(19), 1641–1646 (2011) https://doi.org/10.1016/j.cub.2011.08.031

[47] Tzanetakis, G., Cook, P.: Musical genre classification of audio signals. IEEE Transactions on Speech and Audio Processing **10**(5), 293–302 (2002) https://doi.org/10.1109/TSA.2002.800560

[48] Norman-Haignere, S., Kanwisher, N.G., McDermott, J.H.: Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. Neuron **88**(6), 1281–1296 (2015)

[49] Moeller, S., Yacoub, E., Olman, C.A., Auerbach, E., Strupp, J., Harel, N., Uğurbil, K.: Multiband multislice ge-epi at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fmri. Magnetic resonance in medicine **63**(5), 1144–1153 (2010)

[50] Dale, A.M., Fischl, B., Sereno, M.I.: Cortical surface-based analysis: I. segmentation and surface reconstruction. Neuroimage **9**(2), 179–194 (1999)

[51] Hoerl, A.E., Kennard, R.W.: Ridge regression: Biased estimation for nonorthogonal problems. Technometrics **12**(1), 55–67 (1970)

[52] Destrieux, C., Fischl, B., Dale, A., Halgren, E.: Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. Neuroimage **53**(1), 1–15 (2010)

[53] Zeghidour, N., Teboul, O., Chaumont Quitry, F., Tagliasacchi, M.: Leaf: A learnable frontend for audio classification. International Conference on Learning Representations (2021)

[54] Gemmeke, J.F., Ellis, D.P.W., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 776–780 (2017). https://doi.org/10.1109/ICASSP.2017.7952261

[55] Glasser, M.F., Coalson, T.S., Robinson, E.C., Hacker, C.D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C.F., Jenkinson, M., *et al.*: A multimodal parcellation of human cerebral cortex. Nature **536**(7615), 171–178 (2016)

[56] Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization (2019). https://arxiv.org/abs/1711.05101

Acknowledgements

We would like to thank Aren Jansen for his review of a draft of this paper. SN was supported by MEXT/JSPS KAKENHI JP18H05522 as well as JST CREST JPMJCR18A5 and ERATO JPMJER1801.

Author Contributions Statement

- Initial idea: T. I. Denk, Y. Takagi, C. Frank and S. Nishimoto
- Research design: T. I. Denk, Y. Takagi, A. Agostinelli, T. Nakai, C. Frank and S. Nishimoto
- Data collection: T. Nakai and S. Nishimoto
- Analysis and implementation: T. I. Denk, Y. Takagi, T. Matsuyama and A. Agostinelli
- Writing: T. I. Denk, Y. Takagi, C. Frank and S. Nishimoto
- Reviewing: all authors
- Supervision: C. Frank and S. Nishimoto

Competing Interests Statement

The authors declare no competing interests.

Table 1 Comparison of different music embedding prediction targets (fMRI-to-embedding). For each of the identification accuracy values the standard deviation across five subjects was below 0.016 on the test set; for the correlations it was at most 0.020 on the test set. Identification accuracy and correlation are computed between embeddings of GTZAN audio (computed with the embedding model listed in the respective row) and the embeddings predicted by the regression model. The magnitude of the correlation is strongly influenced by the expressiveness of the chosen embedding. The main takeaway from it is to establish an ordering of the different possible regression targets. To further analyze the behavior, we plot the predicted MuLan embeddings in Supplementary Figure 4 and observe a noisy clustering. An analysis of the regression regularization hyperparameter tuning, which is critical to avoid overfitting, is provided in Section 4.2; more detailed properties of the embeddings in Supplementary Table 1.

Embedding	Ident. accuracy		Correlation	
	Test	Train	Test	Train
SoundStream-avg	0.674	0.764	0.184	0.255
w2v-BERT-avg	0.837	0.941	0.113	0.167
MuLan$^{\text{TEXT}}$	0.817	0.877	0.181	0.245
MuLan$^{\text{MUSIC}}$	0.876	0.992	0.307	0.538

Table 2 Selected ROIs for the Decoding

Subject	Selected Regions
Sub001	ctx_rh_G_temp_sup-Lateral ctx_rh_S_temporal_transverse ctx_lh_S_temporal_transverse ctx_rh_G_temp_sup-Plan_tempo ctx_lh_G_temp_sup-Plan_tempo ctx_rh_G_temp_sup-G_T_transv
Sub002	ctx_lh_Lat_Fis-post ctx_lh_S_temporal_transverse ctx_rh_G_temp_sup-Lateral ctx_rh_S_temporal_transverse ctx_rh_G_temp_sup-G_T_transv ctx_lh_G_temp_sup-G_T_transv
Sub003	ctx_rh_G_temp_sup-Lateral ctx_lh_S_temporal_transverse ctx_lh_Lat_Fis-post ctx_rh_G_temp_sup-Plan_tempo ctx_lh_G_temp_sup-G_T_transv ctx_rh_G_temp_sup-G_T_transv
Sub004	ctx_lh_G_temp_sup-Lateral ctx_rh_G_temp_sup-Lateral ctx_rh_G_temp_sup-G_T_transv ctx_lh_S_circular_insula_inf ctx_lh_Lat_Fis-post ctx_lh_G_temp_sup-G_T_transv
Sub005	ctx_lh_G_temp_sup-Lateral ctx_rh_S_temporal_transverse ctx_lh_S_temporal_transverse ctx_lh_G_temp_sup-G_T_transv ctx_rh_G_temp_sup-Lateral ctx_rh_G_temp_sup-G_T_transv

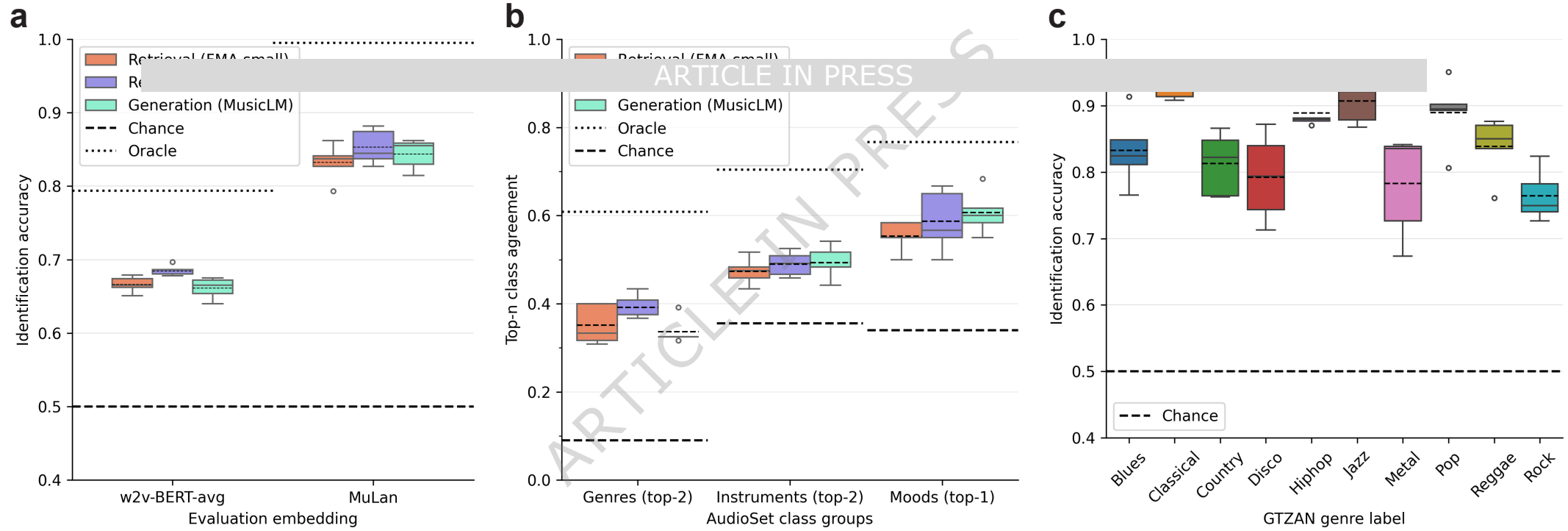
Fig. 1 Main quantitative results of the decoding, i.e., music generation from fMRI signal.
The dashed, horizontal lines (chance) indicate the performance of a random music predictor (sampling from the FMA dataset). The dotted lines (oracle) are the oracle performance, corresponding to the performance achieved by a regressor which would predict exactly the ground truth MuLan embedding of GTZAN. Error bars indicate standard error of the mean across five subjects. **a** Identification accuracy for different evaluation embeddings. Identification accuracy computed between the embeddings of the original stimulus (music) and the embeddings of the generated music. The generated music is more similar to the stimulus it was derived from with respect to high-level embeddings (MuLan) than the low-level w2v-BERT-avg. Differences between generation and retrieval on FMA large (about 106k clips) are marginal, whereas retrieving from FMA small (8k clips) is overall worse. **b** AudioSet top-n class agreement for different groups of AudioSet classes. A list of the classes in each group is provided in Section 2 of Supplementary Information. Generation outperforms retrieval from FMA (both small and large). The worst performance – relative to chance and oracle – is attained on the instrument agreement. **c** Identification accuracy (based on MuLan embeddings of original and generated music) shown separately for each of the GTZAN genres. The model performance is consistent across all genres. Boxes span the 25th–75th percentiles, and the center line marks the median (50th percentile). Whiskers extend to the minimum and maximum values that are not considered outliers—no farther than 1.5 times this range from the 25th and 75th percentiles—and points beyond are outliers. Source data are provided as a Source Data file.

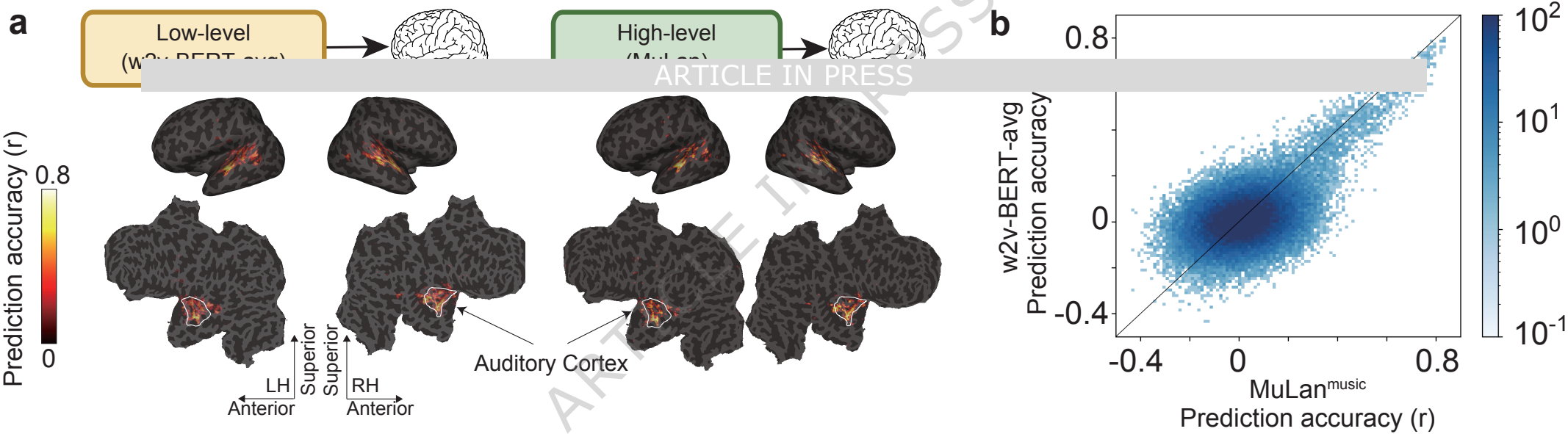
Fig. 2 Comparison between different audio-derived embeddings.
a Prediction performance (measured using Pearson's R) for the voxel-wise encoding model applied to held-out test music on Subject 1, projected onto the inflated (top, lateral and medial views) and flattened cortical surface (bottom, occipital areas are at the center), for both left and right hemispheres. Brain regions with significant accuracy are colored (all colored voxels $p < 0.05$, FDR corrected). White boxes mark anatomically defined auditory cortex. **b** Density plot of the MuLan$^{\text{music}}$ (x-axis) versus w2v-BERT-avg (y-axis) model prediction accuracy. Darker colors indicate a higher number of voxels in the corresponding bin. Source data are provided as a Source Data file.

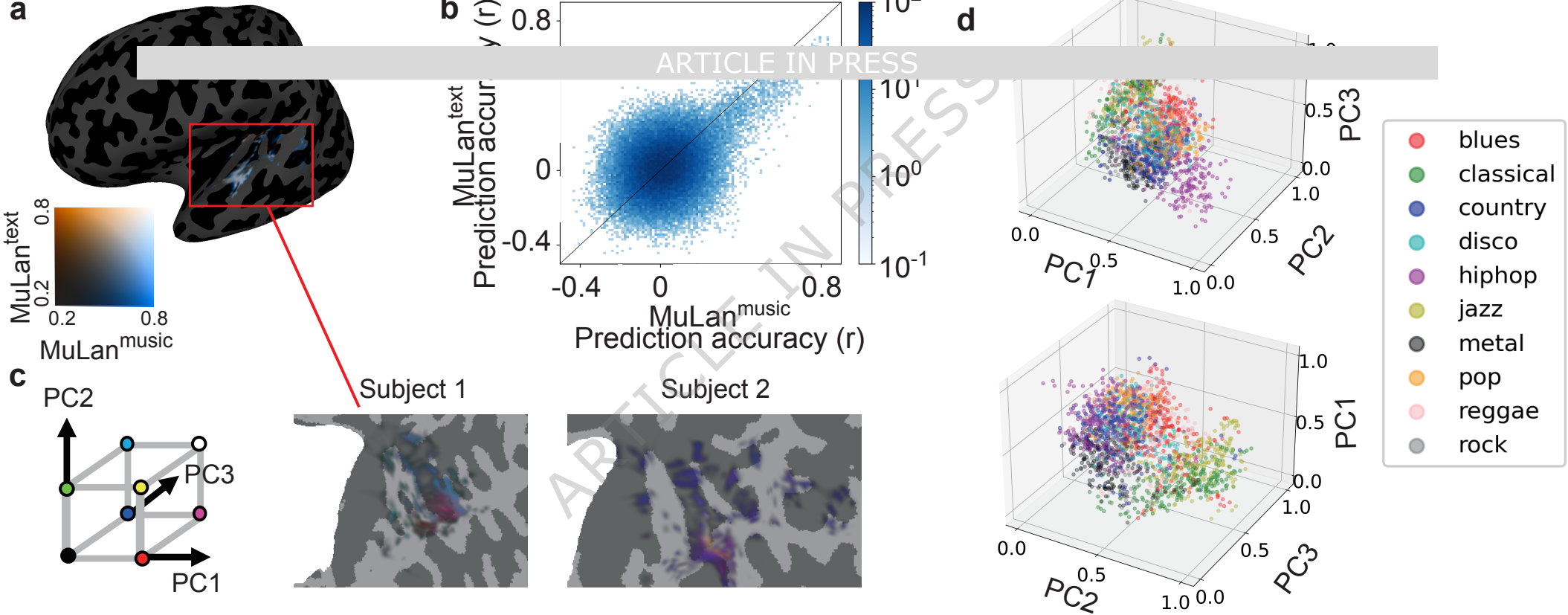
Fig. 3 Comparison between audio- and text-derived MuLan embeddings.
a Prediction performance on subject 1 for $\text{MuLan}^{\text{text}}$ model versus $\text{MuLan}^{\text{music}}$ model. All colored voxels $p < 0.05$, FDR corrected. The area in the red rectangle corresponds to the auditory cortex. **b** Density plot of the $\text{MuLan}^{\text{music}}$ (x-axis) versus $\text{MuLan}^{\text{text}}$ (y-axis) model prediction accuracy. **c** Principal components analysis reveals gradient from medial to lateral axis within auditory cortex. **d** Visualization of the representational relationships between each music projected onto the PC space. The colors represent the genre to which each sample belongs. The top and bottom panels show the same data, but plotted from different angles. Source data are provided as a Source Data file.

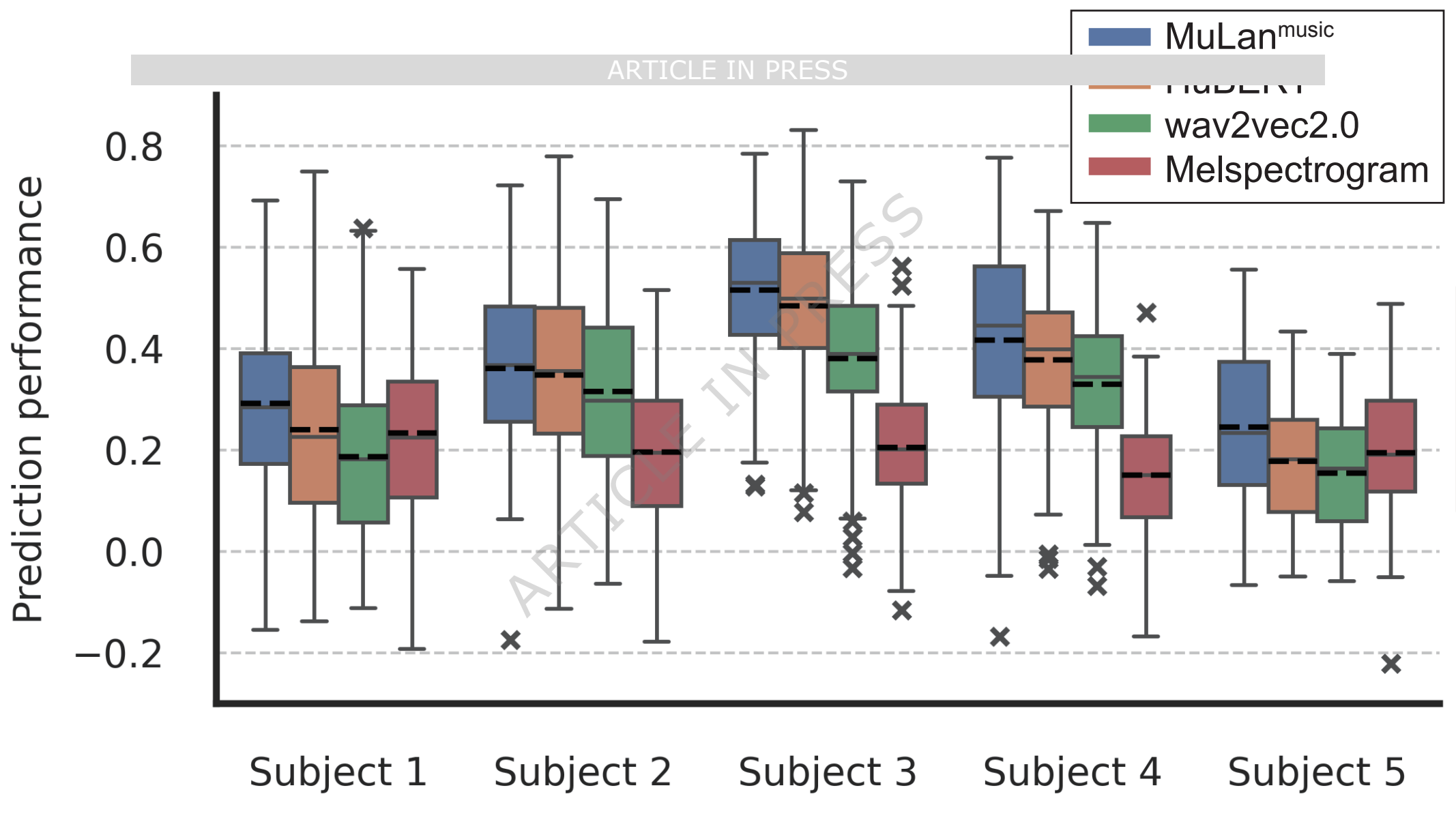
Fig. 4 Comparison with other audio models.
Prediction performance by different models for all subjects. Error bars indicate standard error of the mean across voxels within primary auditory cortex. Boxes span the 25th–75th percentiles, and the center line marks the median (50th percentile). Whiskers extend to the minimum and maximum values that are not considered outliers—no farther than 1.5 times this range from the 25th and 75th percentiles—and points beyond are outliers. Source data are provided as a Source Data file.

Fig. 5 Generalization beyond music genre.
a Comparison of identification accuracy between the models in Figure 1 (full) and those models trained with one-genre-out ablation data and tested on the ablated genre (ablation). Boxes span the 25th–75th percentiles, and the center line marks the median (50th percentile). Whiskers extend to the minimum and maximum values that are not considered outliers—no farther than 1.5 times this range from the 25th and 75th percentiles—and points beyond are outliers. **b** Comparing prediction performance obtained by MuLan$^{\text{music}}$ and genre model on Subject 1. Only the voxels with $R > 0.4$ are colored. Source data are provided as a Source Data file.

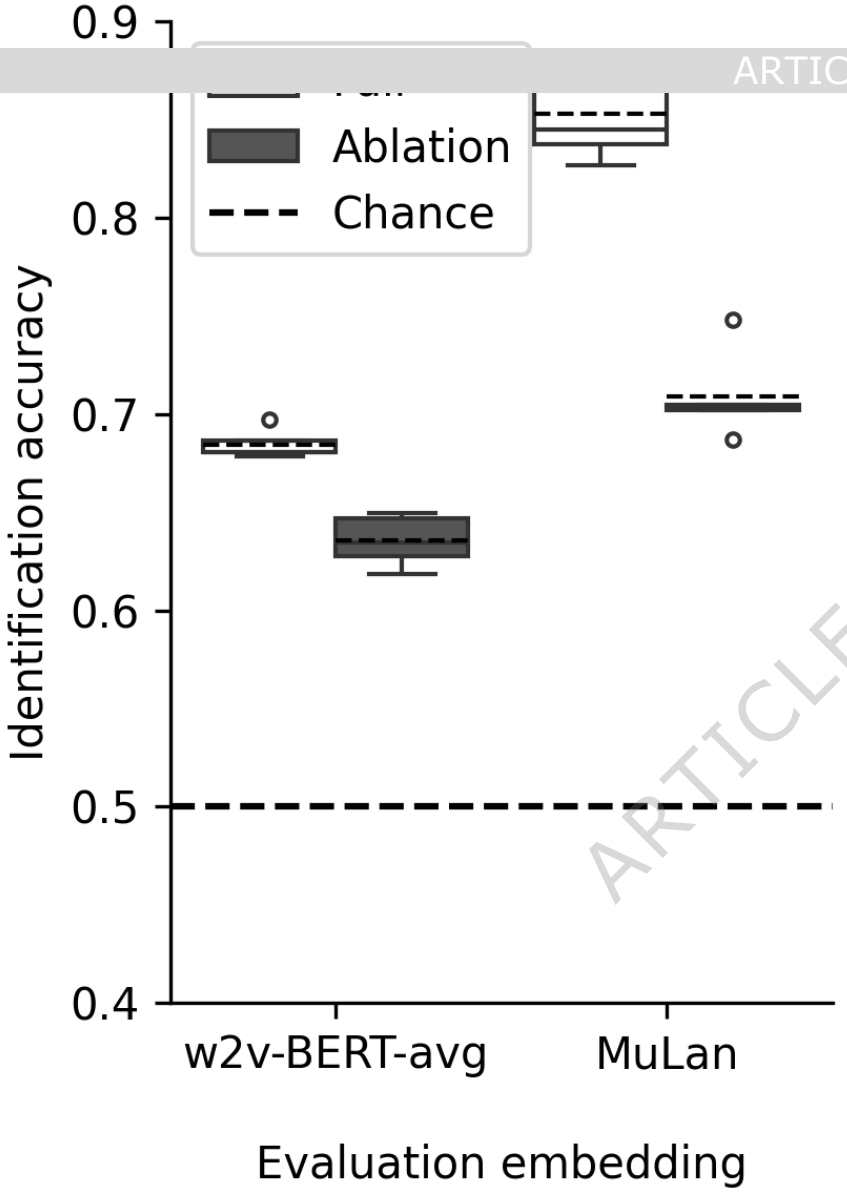








a



b