scientific reports

OPEN

Predicting artificial neural network representations to learn recognition model for music identification from brain recordings

Taketo Akama[1]✉, Zhuohao Zhang[1], Pengcheng Li[1], Kotaro Hongo[1], Shun Minamikawa[1] & Natalia Polouliakh[1]

Recent studies have demonstrated that the representations of artificial neural networks (ANNs) can exhibit notable similarities to cortical representations when subjected to identical auditory sensory inputs. In these studies, the ability to predict cortical representations is probed by regressing from ANN representations to cortical representations. Building upon this concept, our approach reverses the direction of prediction: we utilize ANN representations as a supervisory signal to train recognition models using noisy brain recordings obtained through non-invasive measurements. Specifically, we focus on constructing a recognition model for music identification, where electroencephalography (EEG) brain recordings collected during music listening serve as input. By training an EEG recognition model to predict ANN representations-representations associated with music identification-we observed a significant improvement in classification accuracy. This study introduces a novel approach to developing recognition models for brain recordings in response to external auditory stimuli. It holds promise for advancing brain-computer interfaces (BCI), neural decoding techniques, and our understanding of music cognition. Furthermore, it provides new insights into the relationship between auditory brain activity and ANN representations.

Neural decoding in the audio domain aims to extract or reconstruct auditory information from brain recordings obtained when the audio stimulus is processed by the brain[1–6]. This technique provides insights into the neural localization and dynamics relevant to auditory stimuli and their features[5,7,8]. Furthermore, it enables the exploration of individual differences in cognition, such as those between experts and non-experts or native and non-native listeners[8,9]. Neural decoding in the audio domain also has various research and industrial applications as part of brain-computer interface (BCI) systems, including speech production[2–4,10], music identification[6,11–19], and music reconstruction[1,5,8,20].

Methods for measuring brain responses to auditory stimuli include invasive techniques such as electrocorticography (ECoG) and non-invasive approaches such as functional magnetic resonance imaging (fMRI), electroencephalography (EEG), magnetoencephalography (MEG), and near-infrared spectroscopy (NIRS)[21]. Among these, EEG is particularly suited for high-temporal-resolution time-series data due to its high sampling rate[21,22]. Its affordability, ease of use, and portability suggest significant potential for widespread adoption[21]. However, EEG faces persistent challenges, including low signal-to-noise ratio and attenuation from deeper brain sources[22,23].

Recently, studies have reported that artificial neural network (ANN) representations exhibit similarities to cortical representations in response to the same auditory stimulus[9,24–28]. These studies explore the feasibility of predicting cortical representations by regressing from ANN representations, thereby probing the alignment between these two types of representations.

Our hypothesis posits that if auditory brain response recordings under ideal conditions resemble the representations of artificial neural networks (ANNs), then the information contained in ANN representations may be useful for complementing incomplete brain response recordings obtained under less-than-ideal conditions. We attempt to validate this hypothesis by treating ANN representations as the target signals to be predicted, training a transformation function that complements the measured brain response information, and observing improvements in the accuracy of neural decoding models (recognition models) utilizing the complemented brain response information. In prior studies evaluating the similarity between brain and ANN representations

[1]Sony Computer Science Laboratories, Inc, Tokyo, Japan. [2]Pengcheng Li and Kotaro Hongo: Work conducted when working as a research assistant. ✉email: taketo.akama@sony.com

in response to auditory stimuli, cortical representations were predicted from ANN representations using small linear models, and a high prediction accuracy was interpreted as evidence of similarity. In contrast, our approach reverses this direction of prediction. Specifically, we train a transformation model to enable recorded auditory brain responses to predict ANN representations. Rather than merely evaluating similarity, our objective is to leverage the similarity for information complementation. For this reason, we employ larger, non-linear models. The ultimate goal is to effectively train auditory neural decoding models (recognition models) even when noisy brain recordings are obtained through non-invasive measurements, by using ANN representations as the target signals to be predicted. Figure 1 illustrates the conceptual framework of our approach.

In this study, we focus on developing a recognition model for music identification using EEG brain recordings collected during music listening. By training a recognition model with EEG inputs to infer representations capable of predicting the ANN representations used for music identification, we observed significant improvements in classification accuracy. Furthermore, the model demonstrated robust learning that was less dependent on the initialization of model weights. The most notable improvement in classification accuracy was observed when assuming a brain response delay of approximately 200 ms to music stimuli. This aligns with recent scientific findings on brain response delays to music stimuli[8], suggesting that the model effectively utilizes the similarity between ANN and brain representations. Additionally, the model exhibited an ideal property where longer durations of EEG input resulted in higher music identification accuracy. We also report on the variations in accuracy based on individual differences and differences in the music itself, offering insights from a neuroscience perspective.

To predict ANN representations, various methods can be considered. In this work, we propose using a contrastive learning approach[29–31]. Unlike conventional contrastive learning, our method incorporates two major techniques. The first technique involves simultaneously solving a classification task on both the brain representation branch and the ANN representation branch when using ANN representations as the supervisory signal. This dual-task setup reduces the learning of irrelevant features, generates supervisory signals specific to the target task, and focuses the model on learning features essential for the current task. The second technique is that the ANN representation branch is not trained to align with the brain representation branch. Aligning the ANN representation branch too closely with the brain branch risks diminishing the discriminative capability of the ANN representations for the target task due to the noise inherent in brain recordings. In practice, we confirmed through experiments that implementing this second technique improves music identification accuracy.

This study proposes a novel approach to recognition models for brain recordings in response to music auditory stimuli. It holds promise for advancing the understanding of cognitive mechanisms through neural decoding, contributing to the development of brain-computer interfaces (BCI), and offering valuable insights into the relationship between human music perception and ANN processing of music.

Results

This study utilizes the NMED-T dataset, comprising EEG data collected from 20 subjects as they listened to 10 unique songs[32]. We approach the task as a 10-class classification problem. The input data consists of EEG signals, while the output is the corresponding song ID. Given the 10 possible classes (one for each song), the accuracy of the chance level for this classification task is 0.1.

We begin by optimizing the strength of predicting ANN representation to enhance model performance, followed by an assessment of robustness across various random seed values. We also adjust the model to consider delays between audio onset and participant perception. A comparison between 1D and 2D CNN models, as well as learning techniques , helps identify the optimal configuration. After selecting the best model, we benchmark its performance against prior studies. Further evaluations are conducted on extended duration of EEG to assess flexibility. Finally, we examine model performance on individual songs and explore differences across participants, offering insights from both computational and neuroscientific perspectives.
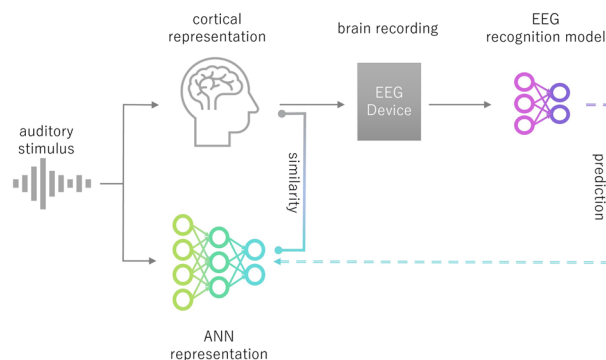
Fig. 1. Conceptual framework of our approach: predicting ANN representation to learn auditory EEG recognition model. When constructing a recognition model that uses brain recordings obtained in response to an auditory stimulus as input, the model is trained to predict the ANN representation obtained by inputting the same auditory stimulus into the ANN. This framework improves the performance of the recognition model by effectively utilizing the findings that cortical representations and ANN representations resemble each other when the same auditory stimulus is inputted.

Preliminary model testing

We adopt the SampleCNN architecture as the base 1D CNN model[33] both for audio and EEG models and extend it with two projectors: one for the classification task and another for the contrastive learning task. To refine our model's performance, parameter tuning was conducted focusing on optimizing the strength of predicting ANN representation via the PredANN loss weight (see "Methods" section for more detail). We initiated this process by comparing weights of 0.01, 0.05, and 0.1 to clarify the optimal PredANN loss weight, and the weight of 0.05 was identified as the optimal value for our model. To ensure robustness and precision in our findings, we further scrutinized the PredANN loss weights ranging from 0.03 to 0.07 at intervals of 0.01. The initial model was tested based on a refined optimal configuration comprising a PredANN loss weight of 0.05 and a seed of 42. The seed value determines the initial state of the model's parameters, and different seeds can lead to different initialization of parameters, impacting the reproducibility and the performance of the model. Seed 42 is a commonly utilized value in the scientific community due to its historical prevalence in computer science. Under this configuration, the model achieved a classification accuracy of 0.482. For comparative purposes, a baseline model with a PredANN loss weight of 0 was also evaluated, yielding a classification accuracy of 0.474. McNemar's test, which is a statistical method used on paired nominal data to assess whether two models differ significantly in their predictions on the same dataset, indicated a statistically significant difference between the models ($p = 1.70 \times 10^{-8}$), suggesting that predicting ANN representation plays a critical role in the model's performance.

Robustness testing

To assess the robustness of our model against variations in random seed values, we conducted a series of experiments. While the primary evaluation was based on seed value 42, we broadened our investigation to include multiple seeds: 0, 1, and 2. This was executed for both our optimal PredANN loss weight of 0.05 and the baseline weight of 0. The objective was to discern whether our proposed model consistently outperforms the baseline across varied initializations. Results from the experiments were surprising for seed values 0 and 1. For these seeds, the baseline model's classification accuracy displayed a degree of unpredictability. In contrast, our proposed model exhibited more consistent and superior performance, as shown in Table 1. These outcomes indicate that our model not only excels at feature extraction but may also possess the capacity to process complex datasets and model architectures. We conducted McNemar's test to compare the performance of our model with the baseline model. The results indicated significant differences for seed 0 ($p = 3.44 \times 10^{-250}$), seed 1 ($p = 8.06 \times 10^{-311}$), and seed 42 ($p = 1.70 \times 10^{-8}$) whereas seed 2 ($p = 0.644$) did not reveal a statistically significant difference. On average, our model achieved a 14.1% improvement in classification accuracy compared to the baseline across all seeds, highlighting the practical significance of the proposed approach.

To provide a comprehensive evaluation, we considered two primary metrics: the maximum classification accuracy across all seeds and the average accuracy across the seed values. Both these metrics underscored the superior performance of our proposed model relative to the baseline.

Incorporating time-delay

We hypothesized that introducing relative delays between music auditory signal and its EEG encoding may enhance EEG classification accuracy. This was based on previous findings that suggested optimal EEG encoding of music occurs approximately 200 ms post-stimulus onset[8,34]. To explore this, we introduced varied latencies between EEG and the music stimulus to evaluate model performance.

In our initial analysis, we examined delay intervals of 80 ms, 160 ms, 320 ms, and 640 ms, using seed values of 0, 1, and 2 to assess robustness. The model's performance for these intervals is shown in the subsequent table, where we evaluated overall efficacy using both the **maximum** and **average** values across all seeds. This approach allowed us to capture both peak performance and consistent performance trends across different conditions. No clear peak performance emerged among these four delays. Thus, we further tested intermediate values, specifically 240 ms (between 160 ms and 320 ms) and 480 ms (between 320 ms and 640 ms). Results indicated that 240 ms yielded superior accuracy in both maximum and average metrics. To further refine the optimal delay, we examined 200 ms-the midpoint between 160 ms and 240 ms-and found it provided the highest accuracy, suggesting it as the optimal delay, which is consistent with previous studies. Table 2 and Fig. 2 summarize classification accuracies across the delay intervals, with the 200 ms model demonstrating the best overall performance.

We also performed McNemar's test to compare the results between a delay of 0 ms and a delay of 200 ms across all seeds. The test revealed significant differences for seed 0 ($p = 7.74 \times 10^{-10}$) and seed 1 ($p = 0.0043$), while no significant difference was found for seed 2 ($p = 0.319$). Incorporating a 200 ms delay led to a 1.7%

PredANN loss weight	Seed					
	0	1	2	42	Maximum	Average
0.05 (our model)	**0.424**	**0.473**	**0.490**	**0.482**	**0.490**	**0.465**
0	0.100	0.100	0.486	0.474	0.486	0.324

Table 1. Accuracy of proposed model for different seeds. With the PredANN loss weight of 0.05, the model achieved significantly higher accuracy for three out of four tested seeds: seed 0 ($p < 0.001$), seed 1 ($p < 0.001$), and seed 42 ($p < 0.001$). Boldface indicates the highest value in each column for ease of comparison.

Delay (ms)	Seed				
	0	1	2	Maximum	Average
0	0.424	0.473	0.490	0.490	0.465
80	0.467	0.483	0.484	0.484	0.478
160	0.459	0.464	0.485	0.485	0.469
200	0.483	0.467	0.495	0.495	0.482
240	0.478	0.464	0.494	0.494	0.479
320	0.460	0.477	0.485	0.485	0.474
480	0.488	0.464	0.473	0.488	0.475
640	0.473	0.446	0.460	0.473	0.459

Table 2. Accuracy of time delay model for different seeds. This table presents the results for different delay settings, with the 200 ms delay yielding the highest accuracy. Statistical significance was observed in comparison to no delay for seed 0 ($p < 0.001$) and seed 1 ($p = 0.0043$), whereas seed 2 did not exhibit statistical significance ($p = 0.306$).
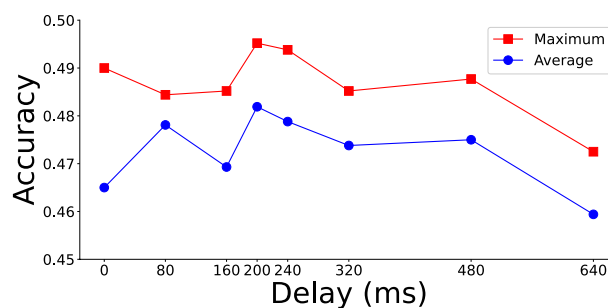
Fig. 2. The accuracy of different delays. The line graph demonstrates the impact of different delay intervals on classification accuracy, aiming to identify the optimal temporal alignment between EEG and auditory stimuli, with the orange line representing the maximum method and the blue line denoting the average method. Both evaluation methods exhibited a peak at a delay of 200 ms. This aligns with prior research, suggesting that this delay corresponds to the typical human auditory reaction time of musical onset.

Model (delay 200 ms)	Seed				
	0	1	2	Maximum	Average
2D CNN, PredANN loss weight 0.05 (our best model)	**0.662**	**0.622**	**0.588**	**0.662**	**0.624**
1D CNN, PredANN loss weight 0.05	0.487	0.465	0.494	0.494	0.482
2D CNN, PredANN loss weight 0	0.537	0.589	0.516	0.589	0.547
2D CNN, PredANN loss weight 0.05, stop-gradient-free	0.648	0.529	0.315	0.648	0.497

Table 3. Accuracy for different models (delay 200 ms). This table presents the results of various CNN comparisons, with the 2D CNN configured with a PredANN loss weight of 0.05 and applying stop-gradient to the music CNN achieving the best performance. This configuration was selected as the optimal model for further analysis. Boldface indicates the highest value in each column for ease of comparison. Statistical significance is reported in the main text.

improvement in classification accuracy, highlighting the practical relevance of temporal alignment in the proposed framework.

2D- versus 1D-CNN performance and stop-gradient effects

After validating the 1D CNN model, we compared the 1D CNN with the 2D CNN. The 2D CNN architecture was selected based on prior research[35]. We tested 200 ms delay with seed values 0, 1, and 2 and results showed that the 2D CNN outperformed the 1D CNN, as shown in Table 3. Additionally, we compared the baseline models (2D CNN with PredANN loss weight 0) and obtained the expected results. As we mentioned in the Introduction, in our main proposed model, the ANN representation model (music model) is not trained to align with the brain representation model (EEG model). Specifically, by applying the stop-gradient method, we ensure that the gradients of the PredANN loss, the purpose of which is to predict ANN representation from the EEG representation, are used to update the EEG model, but not the music model (see "Methods" section for more
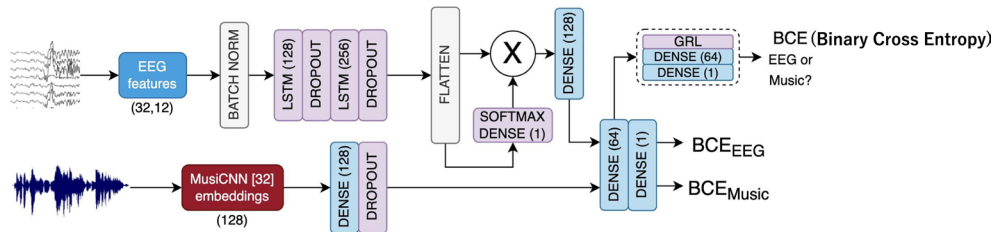
Fig. 3. Previous model structure. The network proposed in the previous study[36] utilized a common layer to align EEG and music modalities, along with a Gradient Reversal Layer (GRL) and Binary Cross-Entropy (BCE) loss for domain adaptation.

Method	Seed 0	1	2
Previous study	0.178	0.144	0.155
Our best model	0.662	0.622	0.588

Table 4. Comparison with previous study. This table presents the comparison between the model in the previous study and our proposed model. Our model outperformed the previous approach across all seeds, demonstrating statistically significant improvements: seed 0 ($p < 0.001$), seed 1 ($p = 0.0016$), and seed 2 ($p < 0.001$).

detail). To validate the effectiveness of our model, we also compare the results with and without the use of the stop-gradient method (stop-gradient-free model).

To compare the performance between the 1D CNN and 2D CNN models, we conducted McNemar's test between the 1D CNN (PredANN loss weight 0.05) and the 2D CNN (PredANN loss weight 0.05). Additionally, we compared the 2D CNN (PredANN loss weight 0.05) with the baseline 2D CNN (PredANN loss weight 0) and the stop-gradient-free 2D CNN (PredANN loss weight 0.05). The results showed that for all seeds (0, 1, and 2), the p-values for the comparison between 1D CNN and 2D CNN were less than 0.001 (seed 1: $p = 1.15 \times 10^{-140}$, seed 2: $p = 4.28 \times 10^{-143}$, seed 3: $p = 4.58 \times 10^{-151}$), indicating significant differences, with the 2D CNN performing considerably better than the 1D CNN. When comparing the 2D CNN (PredANN loss weight 0.05) with the baseline 2D CNN (PredANN loss weight 0), significant differences were observed for seed 0 ($p = 0.035$) and seed 2 ($p = 0.0013$), but no significant difference was found for seed 1 ($p = 0.037$). Lastly, in the comparison between the 2D CNN (PredANN loss weight 0.05) and the stop-gradient-free 2D CNN (PredANN loss weight 0.05), all seeds (0, 1, and 2) yielded p-values of less than 0.001 (seed 1: $p = 2.68 \times 10^{-14}$, seed 2: $p = 1.37 \times 10^{-246}$, seed 3: $p = 4.95 \times 10^{-9}$). Our best model outperformed the 1D CNN by 14.2%, the 2D CNN without cross-modal weighting by 7.7%, and the 2D CNN without stop-gradient by 12.7% in classification accuracy.

Previous study comparison

Subsequently, we compared the Gradient Reversal Layer (GRL) method from previous studies[36]. The GRL enables adversarial training by reversing gradients during backpropagation, encouraging the model to learn features invariant to domain differences. We retained our own 2D CNN model for feature extraction but modified the following layers and loss functions to be consistent with the prior research (Fig. 3). The aim was to compare our method with their GRL method. The results are shown in Table 4. Although the method from previous studies achieved an accuracy above the chance level (10%), our proposed method demonstrated superior performance in the 10-class classification task. To evaluate model performance, we conducted McNemar's test, which revealed statistically significant differences for seed 0 ($p < 4.94 \times 10^{-324}$), seed 1 ($p < 4.94 \times 10^{-324}$), and seed 2 ($p < 4.94 \times 10^{-324}$). These results suggest that our model performs significantly better on the current task. On average, the proposed model achieved a 46.5% improvement in classification accuracy over the model from the previous study across three random seeds, demonstrating substantial performance gains.
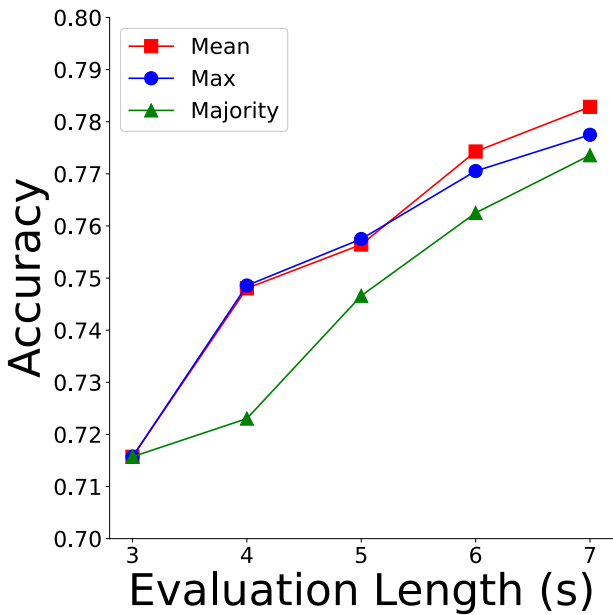
Different evaluation length

We subsequently explored classifying EEG with durations longer than 3 s. By applying overlapping 3-s sliding windows with a 1-s stride, we predicted each window separately to obtain results for the entire duration. Our best-performing model (2D CNN with PredANN loss weight 0.05, delay 200 ms, seed 0) was evaluated across a duration of 3–7 s using three methods: mean (average prediction scores of all windows), max (highest score among windows), and majority (most frequent prediction among windows). Table 5 presents these evaluation results indicating that the mean method achieved the best performance.

We conducted McNemar's test to compare the 3-s and 7-s evaluation periods across all seeds (0, 1, and 2) and all methods (mean, max, and majority). The results showed that for both the max and mean methods, all seeds (0, 1, and 2) yielded p-values of less than 0.001 (max: $p = 7.02 \times 10^{-10}$ for seed 0, $p = 1.38 \times 10^{-9}$ for seed

Length (s)	Method		
	Mean	Max	Majority
3	0.716	0.716	0.716
4	0.748	0.749	0.723
5	0.756	0.758	0.747
6	0.774	0.771	0.763
7	0.783	0.778	0.774

Table 5. Accuracy for different evaluation lengths. This table presents the results for different evaluation lengths, showing that accuracy improves as the evaluation length increases.

Fig. 4. The accuracy of longer evaluation length for three methods. The line graph illustrates the variation in accuracy across different evaluation lengths, with the orange line representing the mean method, the blue line denoting the max method, and the green line corresponding to the majority method. All three evaluation methods exhibit a consistent increase in accuracy as the evaluation length extends, highlighting the robustness and flexibility of the proposed model in handling longer evaluation sequences. Among the three strategies, the mean method appears to be the most promising for future applications, given its consistently high performance across different evaluation lengths.

1, $p = 0.0001$ for seed 2; mean: $p = 1.72 \times 10^{-12}$ for seed 0, $p = 7.06 \times 10^{-12}$ for seed 1, $p = 5.03 \times 10^{-5}$ for seed 2), indicating significant differences between the evaluation lengths. However, for the majority method, only seed 1 demonstrated a significant difference ($p = 0.0074$), while no significant differences were observed for seed 0 ($p = 0.819$) and seed 2 ($p = 0.746$). The proposed model yielded accuracy improvements of 6.7%, 6.2%, and 5.8% when using the Mean, Max, and Majority aggregation methods, respectively.

Figure 4 illustrates the results of the three methods. It can be observed that the accuracy of all methods increases with the evaluation duration, demonstrating that our model is capable of classifying durations longer than 3 ss effectively.
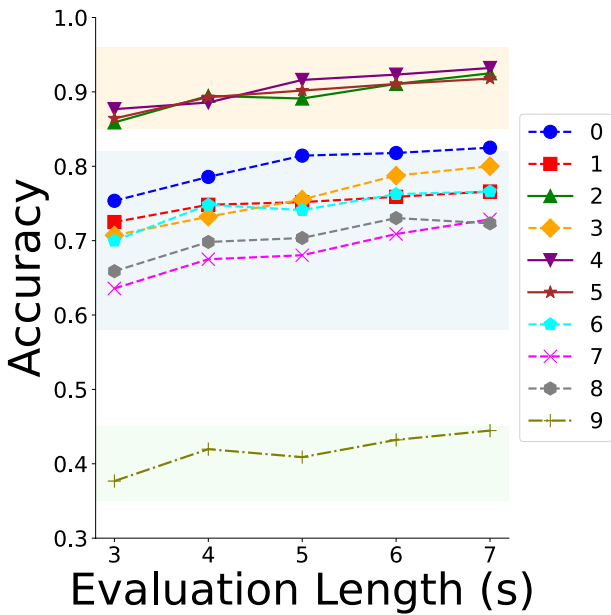
Different songs evaluation

We used the best-performing method mean to evaluate different songs. Each song has unique musical features (such as rhythm and timbre) and we aimed to explore whether these features significantly impact classification accuracy. We hypothesized that unique musical characteristics could influence classification accuracy, making some songs easier to classify. We evaluated durations from 3 to 7 s. Table 6 and Fig. 5 show classification results for different songs at various lengths. Results indicate significant differences in classification accuracy.

All songs were categorized into three distinct groups based on classification accuracy. Songs #2, #4, and #5 demonstrated consistently high accuracy rates, all exceeding 85%. These songs, which are highlighted in the orange group (the solid lines), likely contain unique or easily distinguishable musical elements that facilitate more precise classification. On the other end of the spectrum, song #9 exhibited significantly lower accuracy compared to other songs and is therefore highlighted in the green group (a dot-dash line). This lower performance may indicate the presence of less distinct auditory features, posing a challenge for accurate classification. The

Length (s)	Song_id									
	0	1	2	3	4	5	6	7	8	9
3	0.754	0.725	0.859	0.707	0.877	0.864	0.700	0.636	0.659	0.377
4	0.786	0.748	0.895	0.732	0.886	0.893	0.748	0.675	0.698	0.420
5	0.814	0.752	0.891	0.755	0.916	0.902	0.741	0.680	0.704	0.409
6	0.818	0.759	0.911	0.788	0.923	0.911	0.763	0.709	0.730	0.432
7	0.825	0.766	0.925	0.800	0.932	0.918	0.766	0.729	0.723	0.445

Table 6. Accuracies for different songs. This table presents the results for different evaluation lengths across various songs, illustrating that accuracy varies depending on the song.

Fig. 5. Accuracies for different songs grouped into three levels. The graph presents classification results for individual songs to examine how specific musical characteristics influence model performance, with distinct markers representing each song. The songs are categorized into three groups: top, mid, and bottom, reflecting their respective contributions to model accuracy. This categorization suggests that distinctive musical elements may play a critical role in enhancing model performance.

remaining songs, grouped as the middle range and highlighted in blue (the dashed lines), achieved moderate accuracy levels. These songs may possess characteristics that are somewhat distinguishable but not as pronounced as those in the high-accuracy group. This distribution aligns with our hypothesis that songs with unique and prominent musical elements tend to yield higher classification accuracy. The analysis provides insight into how the distinctiveness of musical features can influence accuracy, which is explored in the "Discussion" section.

Individual evaluation
We then investigated individual performance, hypothesizing that personal differences would be evident in the results. For instance, individuals skilled in music might achieve higher accuracy across more songs, while some may struggle to identify features even in easily classifiable songs. Using the mean method, we evaluated durations from 3 to 7 s. Results shown in Table 7 and Fig. 6 support our hypothesis, highlighting significant individual differences in classification accuracy.

To further analyze individual performance across different songs, we categorized subjects into three groups based on accuracy: top, middle, and bottom. We selected one subject from each group for detailed analysis: subject #3 (top), subject #13 (middle), and subject #14 (bottom). As shown in Fig. 7, the high-accuracy subject demonstrated consistent performance across songs with most accuracies above 80%. The middle group subject showed greater variability ranging from 40% to 100%. The bottom group subject displayed the most variability with accuracies as low as 20%. Interestingly, songs #2, #4, and #5 had high accuracy across all groups suggesting that certain song features may outweigh individual abilities (detailed analysis is provided in the "Discussion" section).

	Subject																			
Length	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	19	20	21	23
3 s	0.742	0.896	0.845	0.611	0.711	0.803	0.763	0.629	0.730	0.579	0.654	0.755	0.500	0.595	0.588	0.765	0.818	0.794	0.879	0.826
4 s	0.773	0.929	0.904	0.639	0.721	0.789	0.814	0.612	0.792	0.621	0.711	0.772	0.523	0.619	0.630	0.769	0.864	0.807	0.951	0.891
5 s	0.752	0.948	0.894	0.693	0.729	0.823	0.797	0.633	0.808	0.625	0.721	0.779	0.539	0.636	0.640	0.782	0.851	0.815	0.940	0.894
6 s	0.783	0.948	0.910	0.711	0.721	0.844	0.823	0.646	0.835	0.639	0.746	0.803	0.552	0.633	0.669	0.789	0.890	0.849	0.940	0.917
7 s	0.770	0.955	0.929	0.696	0.757	0.854	0.840	0.680	0.848	0.646	0.764	0.796	0.575	0.670	0.679	0.769	0.870	0.845	0.934	0.920

Table 7. Accuracies for different individuals. This table presents the results for different evaluation lengths across individuals, illustrating that individual differences significantly impact model performance.
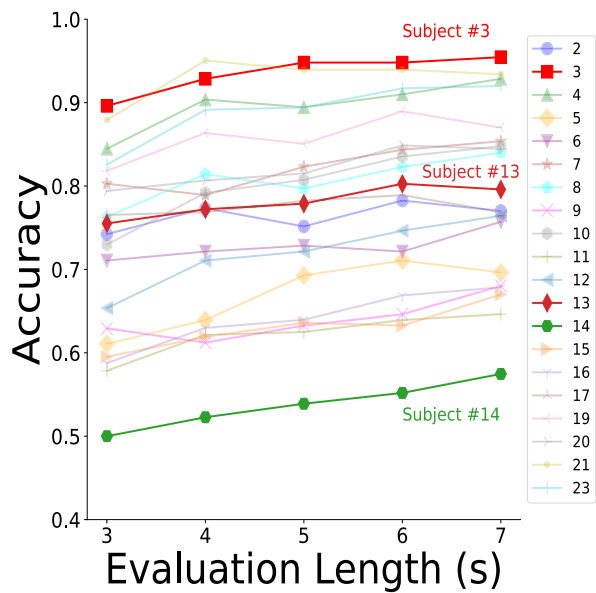
Fig. 6. Accuracies for different individuals. The graph illustrates subject-wise classification accuracy to investigate the effect of individual differences on model performance, with distinct markers representing each subject. Three representative subjects were selected for further analysis: the best-performing subject #3, an average-performing subject #13, and the lowest-performing subject #14. These results highlight the impact of individual differences on model performance.

(a) Subject #3.   **(b)** Subject #13.   **(c)** Subject #14.
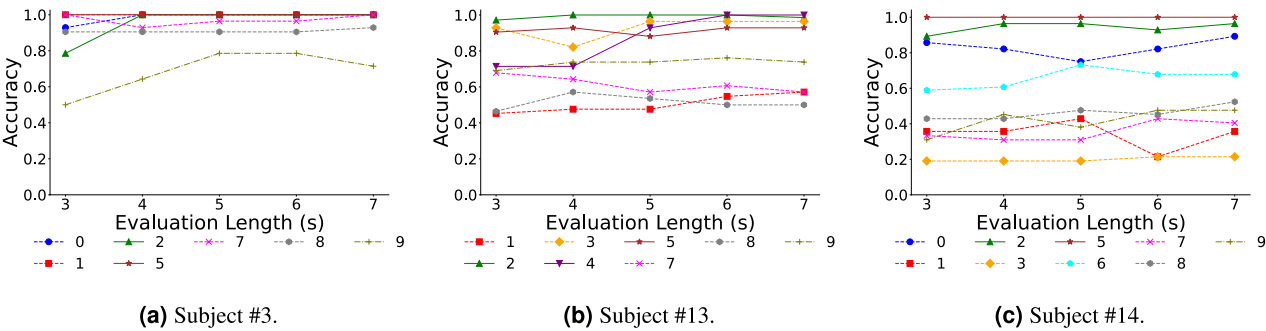
Fig. 7. Individual results for (**a**) top group, (**b**) middle group, (**c**) bottom group. The graph provides a detailed analysis of song-wise classification accuracy for three representative subjects to explore how individual differences interact with specific musical elements, with distinct markers representing different songs. The markers are consistent with those used in Fig. 5. (**a**) Subject #3: High accuracy was achieved across nearly all songs, indicating robust performance. (**b**) Subject #13: Songs in the top group yielded higher accuracy, whereas other songs demonstrated comparatively lower accuracy levels. (**c**) Subject #14: A more diverse range of accuracies was observed, reflecting greater variability in performance across songs. This analysis underscores the varying impact of musical elements and individual differences on model accuracy .

Discussion

Modeling the relationship between auditory stimuli and EEG

Various studies have explored methods to model the relationship between audio stimuli or their features and brain recordings through mathematical functions. These approaches can be broadly categorized into regression, generation, and embedding into a common space. The objectives of these studies vary, including understanding brain function, retrieval, decoding, and elucidating the relationship between ANN and brain representation.

Let us first discuss regression. To capture complex relationships rather than observing ERP (Event-Related Potential), linear models have been used to examine the relationship between auditory stimuli and brain recordings. Modeling from stimuli to recordings is referred to as a forward model (temporal response functions; TRF), while the reverse direction from recordings to stimuli is referred to as a backward model (backward temporal response functions; bTRF)[37]. Forward models have been used to investigate what aspects of stimuli are encoded in EEG. For instance, regression from auditory stimuli or their features to brain recordings has been employed to identify which audio features are important for brain responses[5,8]. Similarly, studies have shown that ANN audio representations resemble brain representations by linearly regressing from ANN representation to brain recordings and analyzing their correlations[9,24–26].

Next, regarding generation, some studies aim to generate realistic music audio directly from brain recordings[1,38,39]. Unlike regression, these approaches often assume a one-to-many mapping for music generation and focus on outputting the audio signal itself rather than its features, prioritizing realism over interpretability.

Finally, methods categorized under embedding into a common space aim to learn a shared space between stimuli and brain recordings for tasks such as retrieval or stimuli reconstruction[10,40]. While our study is closer to this category, it also simultaneously addresses a target classification task. Our objectives include decoding, understanding brain function, and elucidating the relationship between ANN representation and brain representation.

The above-mentioned prior studies, however, were not designed for downstream tasks (e.g., classification) based on representation learning for brain recordings, as is the focus of our study. Although not learning a one-to-one relationship between auditory stimuli and EEG as our study does, the closest concept is found in a study that uses domain adaptation for emotion classification, where the distributional information of music audio representations is reflected in that of EEG representations[41]. In our experiments, we demonstrated the superiority of our approach over this method. By leveraging the one-to-one relationship, our method provides a significantly richer supervisory signal, which we believe underpins the advantage of our proposed approach.

Another interpretation of our framework

As stated in the Introduction, the framework proposed in this study leverages the similarity between auditory cortical representations and ANN representations in a reverse manner. Specifically, it aims to effectively extract essential information from recordings of auditory cortical representations, even when noise is present.

Here, we offer an alternative interpretation of what our framework accomplishes. Auditory brain responses can be understood as signals transformed from auditory stimuli into neural activity. Indeed, methods such as temporal response functions (TRF) aim to approximate this transformation through a mathematical function relationship, thereby facilitating an understanding of the connection between auditory stimuli and their brain responses[37]. From this perspective, our proposed method can be interpreted as performing an inverse transformation, akin to backward temporal response functions (bTRF), where brain responses are transformed back to stimuli. This process attempts to recover stimuli-related information and assists in making inferences about the stimuli.

However, instead of performing an inverse transformation into the data space of the auditory stimuli, our framework transforms the brain responses into the feature space of ANNs. This allows us to obtain high-level feature-based supervisory signals related to the stimuli while avoiding the challenging problem of detailed low-level reconstructions of stimuli that are not directly relevant to the task at hand.

Previous paper comparison

The key distinction between our approach and that of Avramidis et al.[36] lies in the alignment method. While their model employs a 'set-to-set' alignment by integrating music and EEG features into a common layer and distribution matching, our model utilizes a 'point-to-point' alignment by predicting ANN representation using contrastive learning. Our method aligns individual data points directly rather than aligning the two sets of points across modalities, which we hypothesize is the main driver of our model's superior performance. Our 'point-to-point' approach allows for more precise feature mapping, leading to enhanced classification accuracy in multi-class tasks like ours. Moreover, our prediction method with dedicated projection heads effectively captures the nuanced relationships between EEG and music data, avoiding the distortion that can occur when forcing two different modalities to align in the classification head, as seen in the prior method.

In their study, Avramidis et al.[36] set the domain loss weight to 0.1 after determining the optimal weight for their model. Notably, we adjusted the domain loss weight to 0, resulting in a significant performance improvement. Further experimentation with weights of 0.01, 0.05, and 0.001 showed that the best performance was achieved at 0.01, while both 0.05 and 0.001 performed similarly to a weight of 0, all exceeding the performance of the 0.1 setting. These results indicate that the domain loss is not effective in our multi-class song ID classification task.

Longer EEG evaluation

Our analysis indicated that the mean and max methods yielded better classification results compared to the majority method. The max method, by taking the highest prediction value among all windows, minimizes the impact of less distinct windows on the final result. Similarly, the mean method averages prediction values across all windows, enabling the influence of windows with higher accuracy to have a more significant impact.

Window	Predicted values
Window 1	$[\,1.35 \times 10^{-1}, 5.56 \times 10^{-7}, 1.76 \times 10^{-2}, 2.19 \times 10^{-3}, 1.08 \times 10^{-2}, 4.01 \times 10^{-3}, 4.62 \times 10^{-4}, 9.53 \times 10^{-3}, \mathbf{8.19 \times 10^{-1}}, 1.11 \times 10^{-3}\,]$
Window 2	$[\,\mathbf{3.75 \times 10^{-1}}, 1.02 \times 10^{-5}, 8.16 \times 10^{-3}, 2.03 \times 10^{-2}, 1.21 \times 10^{-1}, 1.74 \times 10^{-4}, 3.35 \times 10^{-3}, 1.21 \times 10^{-1}, 3.39 \times 10^{-1}, 1.20 \times 10^{-2}\,]$
Window 3	$[\,7.10 \times 10^{-3}, 3.15 \times 10^{-9}, 4.26 \times 10^{-5}, 2.05 \times 10^{-5}, 1.66 \times 10^{-4}, 3.49 \times 10^{-7}, 1.34 \times 10^{-6}, \mathbf{9.93 \times 10^{-1}}, 9.35 \times 10^{-5}, 2.04 \times 10^{-6}\,]$

Table 8. Predicted values after softmax for each window. This table provides an example of a 5-s evaluation, presenting all three evaluation windows within the 5-s interval. The bold numbers represent the highest prediction value within each window.

Step	Predicted class
Window 1	8
Window 2	0
Window 3	7
Final prediction	8

Table 9. Majority method. The table presents the predicted class for each evaluation window, along with the final prediction aggregated from all windows as the overall prediction for the 5-s period using the majority method.

Step	Predicted values
Max(Window 1, Window 2, Window 3)	$[\,3.75 \times 10^{-1}, 1.02 \times 10^{-5}, 1.76 \times 10^{-2}, 2.03 \times 10^{-2}, 1.21 \times 10^{-1}, 1.74 \times 10^{-4}, 3.35 \times 10^{-3}, \mathbf{9.93 \times 10^{-1}}, 8.19 \times 10^{-1}, 1.20 \times 10^{-2}\,]$
Final prediction	7

Table 10. Max method. The table presents the final predicted values aggregated from each evaluation window using the maximum method, along with the final predicted song number derived from these values for the 5-s evaluation period.

Conversely, the majority method treats each window equally by selecting the most frequently occurring prediction, which may reduce the influence of windows with higher individual accuracy.

As an example, we output a 5-s evaluation period with a true label of 7 divided into three windows. Each window outputs prediction scores for 10 songs (Table 8). The highest score for each window is highlighted in bold. Table 9 shows the majority method. We first determine the final prediction for each window which are 8, 0, and 7 respectively. We then select the most frequent prediction. In this case, the counts are equal so we choose the first value 8 as the final 5-s prediction result. The prediction result is incorrect. However for the max method, as shown in Table 10, we first identify the highest score in each window to obtain the prediction results for all songs. Then we select the highest score among the 10 songs (highlighted in bold). In this example, the score for label 7 in window 3 is significantly high. Using the max method, the influence of window 3 is maximized, resulting in a final prediction of 7, which is correct.

Our results demonstrated that accuracy increased with the length of the evaluation period. We speculate that the different classification accuracies over consecutive windows are due to the difference in informative features, such as unique melodies or rhythms as well as confusing features or EEG noise. This approach offers flexibility, enabling the model to classify not only the 3-s segments used in training but also to adapt seamlessly to longer input lengths without requiring additional training. Our results show that a model trained on 3-s segments generalizes well across various input lengths, often maintaining or even improving classification accuracy. We will discuss about the validation method in more detail and cite related literature in the Different Songs and Subjects section of the "Discussion".

The sliding window method facilitates real-time EEG processing by allowing the model to make continuous evaluations without needing the full signal duration upfront. This capability is particularly valuable for applications requiring immediate feedback, such as brain-computer interfaces, adaptive neurofeedback systems, or real-time auditory processing.

Different songs and subjects

The connection between music and brain activity has been widely acknowledged. Previous studies have demonstrated that musical tempo and rhythm can evoke distinct neural oscillations[6,42,43], while attention to specific musical instruments can lead to unique EEG patterns[44].

Building on these findings, our results suggest that variations in rhythmic and melodic structures, as well as instrumentation, influence EEG responses and subsequent classification performance. We found that songs 2, 4, and 5 consistently achieved high classification accuracy, regardless of the participant's level. We analyzed the characteristics of each song and found that songs with higher accuracy tended to feature electronic elements and unique melodies. We hypothesize that this increased accuracy is due to the presence of synthetic electronic

sounds and distinctive rhythmic patterns, which evoke more pronounced brain responses to these unusual elements. Recent research supports this hypothesis by demonstrating that familiarity with music can lead to varying levels of brain responses. Studies indicate that distinctive elements, such as unique or unfamiliar musical features, tend to evoke more pronounced brainwave activity[45,46]. Notably, song 4 *Lebanese Blonde*, which blends Middle Eastern instruments with electronic effects, creates an exotic and non-traditional melodic scale. This combination likely produces complex neural responses, as the brain processes the unfamiliar and intricate auditory patterns. These distinct neural signatures might make it easier for the classification model to differentiate this song from others, contributing to its higher accuracy.

We also examined individual performance and found that certain participants consistently achieved higher accuracy compared to others. This variability may be attributed to individual differences in music perception, where distinct levels of brain activity are elicited based on factors such as musical training, attentional focus, or psychological states[46]. These individual differences likely contribute to the varying neural responses, impacting the accuracy of music-related tasks across participants[47].

Together, these findings indicate that both the intrinsic properties of songs and individual perceptual factors play key roles in EEG-based music classification accuracy.

Limitations and future work

While our results demonstrate promising advancements in EEG-based music identification, the dataset used in this study[32] is relatively limited in scale, comprising recordings from 20 participants and 10 songs. Although the dataset design allows for detailed within-subject and across-song analyses, its modest size may limit the generalization of our findings. Further research using larger and more diverse datasets appears to be necessary to validate the robustness of our approach across different populations, musical genres, and recording conditions.

Methods
Dataset and preprocessing
Dataset information
We used the Naturalistic Music EEG Dataset-Tempo (NMED-T), a public dataset of EEG recordings from 20 participants who listened to 10 commercially available songs[32]. The dataset contains EEG data collected from 128 scalp electrodes during natural music appreciation. EEG signals were originally sampled at 1000 Hz and then downsampled to 125 Hz to reduce computational cost, while preserving relevant frequency content.

Dataset splitting method
To ensure uniformity, we truncated all recordings to the shared maximum duration of 4 minutes. We partitioned these 4-minute segments into 30-s excerpts. The excerpts were divided into training and validation sets with a 75:25 ratio. To preserve the original distribution of songs in the training and validation sets, we applied stratified sampling during dataset splitting. Specifically, we used the *train_test_split* function from the *sklearn.model_selection* module[48], with the *stratify* parameter set to the song labels. For further details, see the provided code file *preprocessing_eegmusic_dataset.py*.

Preprocessing
In this study, following Défossez et al.[49], we employed the *RobustScaler* for normalizing EEG data, followed by a *clamp* operation to ensure data stability. Specifically, the *RobustScaler* uses the median and interquartile range for scaling, which minimizes the influence of outliers on the data distribution. This approach is particularly suitable for EEG data, as EEG signals may contain sporadic noise and artifacts. For further details, see the *normalize_EEG_4* function in the provided code file *preprocessing_eegmusic_dataset.py*. We use two steps to normalize the data. First, the *RobustScaler* is applied individually to each channel, producing standardized data. Subsequently, a *clamp* operation is used to restrict the values within a specified range (±20).
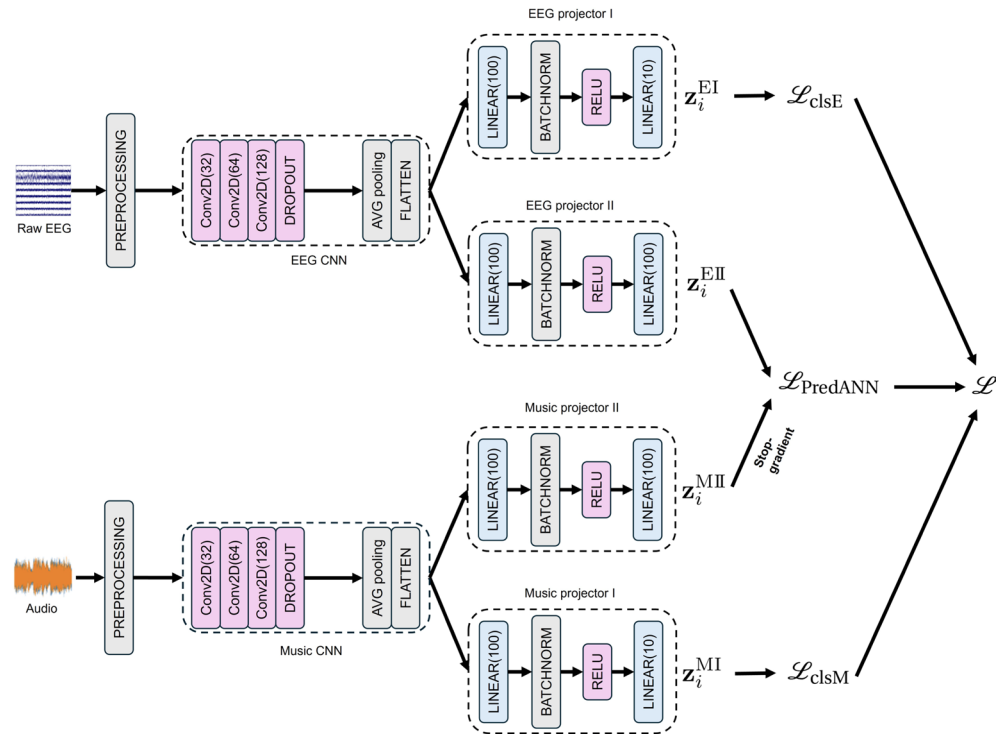
Model
Model architecture and losses
To help the model learn meaningful representations from two different data modalities-EEG and music-we introduce a contrastive loss function called PredANN loss. This loss builds on the InfoNCE loss, a widely used objective in contrastive learning frameworks. Contrastive learning helps the model align EEG and music features by encouraging it to bring together matching pairs and separate unrelated ones. This alignment allows the model to leverage information from music to improve EEG signal classification.

The proposed model consists of two distinct but structurally identical CNN-based encoders: one for processing raw EEG data and another for processing audio data. Each encoder employs 2D CNNs to extract modality-specific features. To maintain consistency in processing both EEG and music data, we applied *padding=1* in all convolutional layers. For EEG data, this ensures that spatial and temporal features are extracted uniformly without reducing the size of spatial and temporal lengths. For music data, *padding=1* allows one-dimensional sequences to be seamlessly processed by 2D convolutions without altering their original length. This unified padding strategy simplifies model design and supports effective feature extraction across modalities (Fig. 8).

The outputs from each encoder are then directed into two separate projectors: Projector I, focused on the classification of 10 songs, and Projector II, dedicated to contrastive learning. The projectors output different types of embeddings for both EEG and music data: song embeddings from Projector I (for classification tasks) and feature embeddings from Projector II (for contrastive learning). The model incorporates both Cross-Entropy (CE) loss for classification and PredANN loss for auxiliary representation learning. The CE loss is computed on the song embeddings to optimize the classification predictions for each class label. Meanwhile,

Fig. 8. The proposed model. The graph illustrates the structure of the proposed model. Two separate 2D CNNs are employed to process music and EEG data independently. The outputs include individual losses for music and EEG, along with a contrastive loss for learning the relationship between the two modalities.

the PredANN loss calculates the similarity between feature embeddings from both EEG and music encoders, encouraging shared feature space alignment between the two modalities. We apply the stop-gradient operation to the music feature embeddings, preventing gradients from backpropagating through the music encoder in the contrastive task. The final loss function is a weighted combination of three components: the EEG classification loss, the music classification loss, and the PredANN loss. This cumulative loss function drives joint optimization, allowing the model to simultaneously learn discriminative features for classification and representations for contrastive learning. Formally, let $\{z_i^{EI}\}_{i=0}^{B}$, $\{z_i^{EII}\}_{i=0}^{B}$, $\{z_i^{MI}\}_{i=0}^{B}$, and $\{z_i^{MII}\}_{i=0}^{B}$ be outputs of EEG projector I, EEG projector II, Music projector I, and Music projector II, respectively, in a mini-batch of size B when training the model. The classification losses for EEG and music are defined as

$$\mathscr{L}_{\text{clsE}} := \sum_{i=0}^{B} \text{CE}\left(z_i^{EI}, c_i^{E}\right) \quad \text{and} \quad \mathscr{L}_{\text{clsM}} := \sum_{i=0}^{B} \text{CE}\left(z_i^{MI}, c_i^{M}\right), \tag{1}$$

respectively, where CE denotes cross entropy loss, and c_i^{E} and c_i^{M} denote classification labels of EEG and music, respectively. The PredANN loss is defined as

$$\mathscr{L}_{\text{PredANN}} := -\sum_{i=0}^{B} \left(\log \frac{\exp\left(\text{sim}\left(\text{sg}(z_i^{MII}), z_i^{EII}\right)/\tau\right)}{\sum_{j=0}^{B} \exp\left(\text{sim}\left(\text{sg}(z_i^{MII}), z_j^{EII}\right)/\tau\right)} + \log \frac{\exp\left(\text{sim}\left(\text{sg}(z_i^{MII}), z_i^{EII}\right)/\tau\right)}{\sum_{j=0}^{B} \exp\left(\text{sim}\left(\text{sg}(z_j^{MII}), z_i^{EII}\right)/\tau\right)} \right), \tag{2}$$

where $\text{sim}(\cdot, \cdot)$, $\text{sg}(\cdot)$, and τ denote cosine similarity, stop-gradient operation, and the temperature parameter, respectively. The PredANN loss is based on the maximization of mutual information between EEG and music embeddings, rather than the maximization of likelihood of conditional probabilities. In practice, we use InfoNCE loss that approximates the negative mutual information, proposed in the context of Contrastive Predictive Coding (CPC) of speech or image[29]. The use of InfoNCE loss between different modalities was first explored in ConVIRT in image and text-domain[30] and later popularized by CLIP[31]. Unlike these prior works, our PredANN loss introduces stop-gradient operations $\text{sg}(\cdot)$ and applied to the EEG and music domain. The final loss function is

$$\mathscr{L} := \mathscr{L}_{\text{clsE}} + \mathscr{L}_{\text{clsM}} + \lambda \mathscr{L}_{\text{PredANN}}. \tag{3}$$

Here is another intuitive explanation of our model. It utilizes separate encoders for EEG and music data, allowing each encoder to specialize in extracting features specific to its modality for the classification task. Rather than

imposing a strict alignment between modalities, we employ a PredANN loss after branching off from the main network that leads to the classification head, preserving the unique characteristics of each modality. Moreover, rather than the 'set-to-set' alignment demonstrated in the prior study[36], 'point-to-point' alignment ensures that the two modalities are aligned at a feature level without distorting the original data representations used for classification (see the "Discussion": Previous Paper Comparison section). This alignment strategy demonstrates significant potential for enhancing EEG classification through the incorporation of music data only during training.

Model training and evaluation
We conducted our training using 6000 epochs to ensure model convergence. For data extraction during training, we implemented a stride of 200, meaning that during each iteration, data points were extracted every 200 steps, which allowed for efficient use of computational resources by reducing the amount of data processed at each training epoch. However, for evaluation, we used a stride of 1 to assess the model's performance on all available data points without omission, enabling a more precise and detailed measurement of accuracy. Although the training stride was coarser, we employed a sliding window and random clipping approach throughout the 6000 training epochs. This data augmentation strategy ensures that each segment of data is likely encountered multiple times during training. Therefore, we consider it reasonable to evaluate the model using a stride of 1 during validation without introducing a distributional mismatch.

To evaluate the effectiveness of our models, we conducted McNemar's test on each of the seeds. For each prediction, a correct result was marked as 1 and an incorrect result as 0, producing an array of binary outcomes for each model. McNemar's test was then applied to compare the binary arrays from two models, enabling us to statistically assess whether the differences in model predictions were significant. This approach provides a rigorous statistical framework for evaluating model performance beyond mere accuracy, offering clearer insight into the models' internal behavior.

Conclusion

In this study, we proposed a method for training auditory EEG recognition models by leveraging the similarity between cortical and ANN representations in response to the same auditory stimuli. The effectiveness of this approach was demonstrated in the music identification task, where performance was enhanced by complementing essential information in EEG recordings through training the recognition model to predict ANN representations.

Our experiments showed significant performance improvements for both 1D CNN and 2D CNN architectures in the music identification task. The model exhibited robust learning and effectively incorporated time delays in brain responses to musical stimuli, aligning with recent findings on temporal delays in auditory neural processing. Moreover, we demonstrated that the model could adapt to longer EEG input sequences, with performance improving as the input length increased.

Additionally, we investigated the effects of individual differences and variations in musical stimuli on identification performance. Notably, even individuals with lower overall accuracy in music identification achieved higher performance with specific musical stimuli, emphasizing that differences in the stimuli were more influential than individual differences.

This study demonstrates improved accuracy in recognition models through a framework based on understanding the relationship between audio stimuli and brain recordings. Consequently, it is expected to contribute to the elucidation of cognitive mechanisms through neural decoding, the advancement of brain-computer interfaces (BCI), and the deepening of insights into the relationship between the human brain and ANN representations.

Data availability

The datasets analyzed during the current study are publicly available[32] and can be accessed via the following link: https://exhibits.stanford.edu/data/catalog/jn859kj8079. Note that in the NMED-T dataset we used, each participant's data does not include all songs. For example, participant 14's data lacks trials for song #4. Additionally, the total number of trials varies between participants. However, the total number of trials for each song is equal, with each song having 560 trials.

Code availability

Upon acceptance, the source code utilized for conducting the experiments will be publicly accessible at https://github.com/JURIUENO11/PredANN.

References

1. Daly, I. Neural decoding of music from the EEG. *Sci. Rep.* **13**, 624. https://doi.org/10.1038/s41598-022-27361-x (2023).
2. Pasley, B. N. et al. Reconstructing speech from human auditory cortex. *PLOS Biol.* **10**, e1001251. https://doi.org/10.1371/journal.pbio.1001251 (2012).
3. Akbari, H., Khalighinejad, B., Herrero, J. L. & Mesgarani, N. Towards reconstructing intelligible speech from the human auditory cortex. *Sci. Rep.* **9**, 874. https://doi.org/10.1038/s41598-018-37359-z (2019).
4. Anumanchipalli, G. K., Chartier, J. & Chang, E. F. Speech synthesis from neural decoding of spoken sentences. *Nature* **568**, 493–498. https://doi.org/10.1038/s41586-019-1119-1 (2019).

5. Bellier, L. et al. Music can be reconstructed from human auditory cortex activity using nonlinear decoding models. *PLOS Biol.* **21**, e3002176. https://doi.org/10.1371/journal.pbio.3002176 (2023).
6. Hoefle, S. et al. Identifying musical pieces from FMRI data using encoding and decoding models. *Sci. Rep.* **8**, 2266. https://doi.org/10.1038/s41598-018-20732-3 (2018).
7. Correia, J. et al. Brain-based translation: FMRI decoding of spoken words in bilinguals reveals language-independent semantic representations in anterior temporal lobe. *J. Neurosci.* **34**, 332–338. https://doi.org/10.1523/JNEUROSCI.1302-13.2014 (2014).
8. Di Liberto, G. M. et al. Cortical encoding of melodic expectations in human temporal cortex. *eLife* **9**, e51784 (2022).
9. Millet, J. *et al.* Toward a realistic model of speech processing in the brain with self-supervised learning. in *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS)* (2022). *Equal contribution.
10. Défossez, A. et al. Decoding speech perception from non-invasive brain recordings. *Nat. Mach. Intell.* **5**, 1097–1107. https://doi.org/10.1038/s42256-023-00714-5 (2023).
11. Marion, G., Di Liberto, G. M. & Shamma, S. A. The music of silence: Part I: Responses to musical imagery encode melodic expectations and acoustics. *J. Neurosci.* **41**, 7435–7448. https://doi.org/10.1523/JNEUROSCI.0183-21.2021 (2021).
12. Foster, C., Dharmaretnam, D., Xu, H., Fyshe, A. & Tzanetakis, G. Decoding music in the human brain using eeg data. In *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*, 1–6, https://doi.org/10.1109/MMSP.2018.8547051 (IEEE, 2018).
13. Sternin, A. *Classifying Music Perception and Imagination Using EEG*. Ph.d. thesis, The University of Western Ontario (2016).
14. Lawhatre, P., Shiraguppi, B. R., Sharma, E., Miyapuram, K. P. & Lomas, D. Classifying songs with EEG. arXiv preprint arXiv:2010.04087 (2020).
15. Stober, S., Sternin, A., Owen, A. M. & Grahn, J. A. Deep feature learning for eeg recordings. arXiv preprint arXiv:1511.04306v4 (2015).
16. Sonawane, D., Miyapuram, K. P., Shiraguppi, B. R. & Lomas, D. J. Guessthemusic: Song identification from electroencephalography response. in *Proceedings of the ACM International Conference Series*, 154–162, https://doi.org/10.48550/arXiv.2009.08793 (2020).
17. Ramirez-Aristizabal, A. G. & Kello, C. Eeg2mel: Reconstructing sound from brain responses to music. arXiv preprint arXiv:2207.13845 (2022).
18. Ramirez-Aristizabal, A. G., Ebrahimpour, M. K. & Kello, C. T. Image-based eeg classification of brain responses to song recordings. arXiv preprint arXiv:2202.03265https://doi.org/10.48550/arxiv.2202.03265 (2022).
19. Pandey, P., Sharma, G., Miyapuram, K. P., Subramanian, R. & Lomas, D. Music identification using brain responses to initial snippets. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1246–1250, https://doi.org/10.1109/ICASSP43922.2022.9747332 (Singapore, Singapore, 2022).
20. Di Liberto, G. M., Marion, G. & Shamma, S. A. Accurate decoding of imagined and heard melodies. *Front. Neurosci.* **15**, 673401. https://doi.org/10.3389/fnins.2021.673401 (2021).
21. Vogel, J. et al. An assistive decision-and-control architecture for force-sensitive hand-arm systems driven by human-machine interfaces. *Int. J. Robot. Res.* **34**, 763–780. https://doi.org/10.1177/0278364914561535 (2015).
22. Daly, I. et al. Electroencephalography reflects the activity of sub-cortical brain regions during approach-withdrawal behaviour while listening to music. *Sci. Rep.* **9**, 9415. https://doi.org/10.1038/s41598-019-45105-2 (2019).
23. Schirrmeister, R. T. et al. Deep learning with convolutional neural networks for EEG decoding and visualization. *Hum. Brain Map.* **38**, 5391–5420. https://doi.org/10.1002/hbm.23730 (2017).
24. Vaidya, A. R., Jain, S. & Huth, A. G. Self-supervised models of audio effectively explain human cortical responses to speech. arXiv preprint arXiv:2205.14252 (2022).
25. Tuckute, G., Feather, J., Boebinger, D. & McDermott, J. H. Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions. *PLOS Biol.* **21**, e3002366. https://doi.org/10.1371/journal.pbio.3002366 (2023).
26. Oota, S. R., Pahwa, K., Marreddy, M., Gupta, M. & Raju, B. S. Neural architecture of speech. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5, https://doi.org/10.1109/ICASSP49357.2023.10096248 (IEEE, 2023).
27. Giordano, B. L., Esposito, M., Valente, G. & Formisano, E. Intermediate acoustic-to-semantic representations link behavioral and neural responses to natural sounds. *Nat. Neurosci.* **26**, 664–672. https://doi.org/10.1038/s41593-023-01285-9 (2023).
28. Oota, S. R. *et al.* Deep neural networks and brain alignment: Brain encoding and decoding (survey). arXiv preprint arXiv:2307.10246 (2024).
29. van den Oord, A., Li, Y. & Vinyals, O. Representation learning with contrastive predictive coding. *ArXiv***abs/1807.03748** (2018).
30. Zhang, Y., Jiang, H., Miura, Y., Manning, C. D. & Langlotz, C. P. Contrastive learning of medical visual representations from paired images and text. in Lipton, Z., Ranganath, R., Sendak, M., Sjoding, M. & Yeung, S. (eds.) *Proceedings of the 7th Machine Learning for Healthcare Conference*, vol. 182 of *Proceedings of Machine Learning Research*, 2–25 (PMLR, 2022).
31. Radford, A. *et al.* Learning transferable visual models from natural language supervision. in Meila, M. & Zhang, T. (eds.) *Proceedings of the 38th International Conference on Machine Learning*, vol. 139 of *Proceedings of Machine Learning Research*, 8748–8763 (PMLR, 2021).
32. Losorelli, S., Nguyen, D. T., Dmochowski, J. P. & Kaneshiro, B. Nmed-t: A tempo-focused dataset of cortical and behavioral responses to naturalistic music. *ISMIR* **3**, 5 (2017).
33. Lee, J., Park, J., Kim, K. L. & Nam, J. Samplecnn: End-to-end deep convolutional neural networks using very small filters for music classification. *Appl. Sci.* **8**, 150. https://doi.org/10.3390/app8010150 (2018).
34. Jagiello, R., Pomper, U., Yoneya, M., Zhao, S. & Chait, M. Rapid brain responses to familiar vs unfamiliar music—an EEG and pupillometry study. *Sci. Rep.* **9**, 15570 (2019).
35. Ramirez-Aristizabal, A., Ebrahimpour, M. & Kello, C. Image-based EEG classification of brain responses to song recordings. arXiv preprint arXiv:2202.03265 (2022).
36. Avramidis, K., Garoufis, C., Zlatintsi, A. & Maragos, P. Enhancing affective representations of music-induced EEG through multimodal supervision and latent domain adaptation. arXiv preprint arXiv:2202.09750 (2022).
37. Crosse, M. J., Di Liberto, G. M., Bednar, A. & Lalor, E. C. The multivariate temporal response function (MTRF) toolbox: A Matlab toolbox for relating neural signals to continuous stimuli. *Front. Hum. Neurosci.* **10**, 604. https://doi.org/10.3389/fnhum.2016.00604 (2016).
38. Postolache, E. *et al.* Naturalistic music decoding from eeg data via latent diffusion models. arXiv preprint arXiv:2405.09062 (2024).
39. Denk, T. I. *et al.* Brain2music: Reconstructing music from human brain activity. arXiv preprint arXiv:2307.11078 (2023).
40. Yu, Y., Beuret, S., Zeng, D. & Oyama, K. Deep learning of human perception in audio event classification. arXiv preprint arXiv:1809.00502https://doi.org/10.48550/arXiv.1809.00502 (2018).
41. Avramidis, K., Garoufis, C., Zlatintsi, A. & Maragos, P. Enhancing affective representations of music-induced eeg through multimodal supervision and latent domain adaptation. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 5, https://doi.org/10.48550/arXiv.2202.09750 (2022). 5 pages, 3 figures.
42. Stober, S., Prätzlich, T. & Mrüller, M. Brain beats: Tempo extraction from EEG data. In *ISMIR*, 276–282 (2016).
43. Rivera-Tello, S., Romo-Vázquez, R., González-Garrido, A. A. & Ramos-Loyo, J. Musical tempo affects EEG spectral dynamics during subsequent time estimation. *Biol. Psychol.* **178**, 108517 (2023).
44. Cantisani, G., Essid, S. & Richard, G. EEG-based decoding of auditory attention to a target instrument in polyphonic music. *2019 IEEE WASPAA* 80–84 (2019).

45. Li, Q., Liu, G., Zhang, Y., Wu, J. & Huang, R. *A functional magnetic resonance study. PsyPost - Psychology News, Neural correlates of musical familiarity* (2024).
46. Ding, K., Li, J., Li, X. & Li, H. Understanding the effect of listening to music, playing music, and singing on brain function: A scoping review of fnirs studies. *Brain Sci.* **14**, 751. https://doi.org/10.3390/brainsci14080751 (2024).
47. Martínez-Molina, N., Mas-Herrero, E., Rodríguez-Fornells, A., Zatorre, R. J. & Marco-Pallares, J. White matter microstructure reflects individual differences in music reward sensitivity. *J. Neurosci.* **36**, 5301–5310. https://doi.org/10.1523/JNEUROSCI.2506-15.2016 (2016).
48. Pedregosa, F. et al. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
49. Défossez, A., Caucheteux, C., Rapin, J., Kabeli, O. & King, J. R. Decoding speech from non-invasive brain recordings. arXiv preprint arXiv:2208.12266 (2022).

Author contributions

T.A. conceptualized and designed the framework, method, and experiment. K.H., P.L., and Z.Z. also designed the method and experiment. Z.Z., K.H., P.L., and T.A. implemented the code and conducted the experiment. Z.Z., T.A., P.L., and N.P. analyzed and discussed the results. Z.Z., T.A., and P.L. wrote the main manuscript and created tables and figures. S.M. and Z.Z. organized the code. T.A., Z.Z., and N.P. reviewed the manuscript. N.P. advised the research and organized the research project.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to T.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.