Alignment of auditory artificial networks with massive individual fMRI brain data leads to generalizable improvements in brain encoding and downstream tasks

Maelle Freteault[1,2,3]*, Maximilien Le Clei[2], Loic Tetrel[2,4], Pierre Bellec[1,2], Nicolas Farrugia[3]

[1] Université de Montréal, Montréal, QC, Canada; [2] Centre de Recherche de L'Institut Universitaire de Gériatrie de Montréal, Montréal, QC, Canada; [3] IMT Atlantique, Brest, France; [4] Kitware Europe, Villeurbanne, France

***For correspondence**: maelle.freteault@gmail.com

Abstract

Artificial neural networks are emerging as key tools to model brain processes associated with sound in auditory neuroscience. Most modelling works fit a single model with brain activity averaged across a group of subjects, ignoring individual-specific features of brain organisation. We investigate here the feasibility of creating personalised auditory artificial neural models directly aligned with individual brain activity. This objective raises major computational challenges, as models have to be trained directly with brain data, which is typically collected at a much smaller scale than data used to train models in the field of artificial intelligence. We aimed to answer two key questions: can brain alignment of auditory models lead to improved brain encoding for novel, previously unseen stimuli? Can brain alignment of auditory models lead to generalisable representations of auditory signals that are useful to solve a variety of complex auditory tasks? To answer these questions, we relied on two massive datasets. First, we used a deep phenotyping dataset from the Courtois neuronal modelling project, where six subjects watched four seasons (36 hours) of the Friends TV series in functional magnetic resonance imaging. Second, we evaluated personalised brain models on a very large battery of downstream tasks called HEAR, where we can rank our models against a collection of recent AI models. Given the moderate size of our neuroimaging dataset, compared with modern AI standards for training, we decided to fine-tune SoundNet, a small and pretrained convolutional neural network featuring about 2.5M parameters. Aligning SoundNet with brain data on three seasons of Friends led to substantial improvement in brain encoding in the fourth season, included but not limited to the auditory and visual cortices. We also observed consistent performance gains on the

HEAR evaluation benchmark. For most tasks, gains were often modest, as our brain-aligned models perform better than SoundNet, and in some cases surpass a few other models. However large gains were observed across subjects for tasks with limited amount of training data, placing brain-aligned models alongside the best performing models regardless of their size. Taken together, our results demonstrate the feasibility of applying AI tools to align artificial neural network representations with individual brain activity during auditory processing, and that this alignment seems particularly beneficial for tasks with limited amount of training data available. Future research is needed to establish whether larger models can be trained as well, with even better performance both for brain encoding and downstream task behaviour, and whether the large gains we observed extend to other downstream tasks with limited training data, especially in the context of few shot learning.

Keywords

Auditory neuroscience, Individual-specific computational models, Artificial neural networks, Downstream generalisation, Deep phenotyping datasets, Functional magnetic resonance imaging (fMRI)

Introduction

Overall objective

Artificial neural networks (ANNs) have emerged as a powerful tool in cognitive neuroscience. Specifically, ANNs trained to solve complex tasks directly from rich data streams, such as natural images, are able to accurately encode brain activity, i.e. predict brain responses directly from the stimulus. A notable observation is that the ANNs which have high performance solving behavioural tasks, e.g. image classification, tend to be the ones which perform best to predict brain activity. This was noted first in vision (Yamins et al., 2014) and rigorously established for language while controlling for model architecture and model capacity (Caucheteux et al., 2023). This result suggests that directly aligning the representations of ANNs with the brain may lead to more generalizable representations and improved performance on novel downstream tasks. Only a few works have explored the possibility of direct alignment of ANNs with brain activity so far, and most of these works have been carried in the field of vision (Seeliger et al., 2021; St-Yves et al., 2023; Lu et al., 2024) and language (Conwell et al., 2021; Schwartz et al., 2019). These brain-alignment works used datasets of limited size, both for brain encoding and downstream tasks. In this work, we explore for the first time the impact of individual brain alignment on an auditory ANN. The study also leverages massive datasets both for testing the generalisation of brain encoding to novel stimuli, with the Courtois NeuroMod dataset (Boyle et al., 2020), and evaluating task performance on a wide range of downstream tasks, with the HEAR benchmark (Turian et al., 2022).

ANNs in audio classification and brain encoding

ANNs trained with deep learning for cognitive neuroscience emerged initially in the field of vision (Schrimpf et al., 2020). CNNs in artificial vision share strong parallels with visual brain processing (Bengio et al., 2013; Krizhevsky et al., 2012). It was found that ANNs trained on complex tasks such as image annotation could predict brain responses to image stimuli with considerable accuracy (Yamins et al., 2014). Building on these foundations, the capability of ANNs for brain encoding has been extended to both language and auditory neuroscience. Kell and collaborators designed an auditory CNN with two branches, tailored for music and speech recognition (Kell et al., 2018). They discovered distinct auditory processing streams in their network, with the primary auditory cortex best predicted by middle layers, and the secondary auditory cortex by late layers. More recently, Giordano and coll. provided further evidence for the intermediary role of the superior temporal gyrus (STG) in auditory processing using 3 different CNNs (Giordano et al., 2023), while Caucheteux and coll. use of

a modified GPT-2 provided new insights into predictive coding theory (Caucheteux et al., 2023).

Brain-augmented learning

Typical brain encoding studies re-use a pretrained network based on a very large collection of sounds, and can feature a very large number of parameters. It is relatively straightforward to apply such large networks for brain encoding, e.g. using Ridge regression from the latent space of the network to brain activity, as it was done for example with BERT (Devlin et al., 2018). Aligning internal representations of models with brain activity, by contrast, requires directly optimising the parameters of the network in order to maximise the quality of brain encoding, which raises a number of computational and conceptual challenges. First, there is clear evidence of substantial inter-individual differences in functional brain organisation (Gordon et al., 2017; Gratton et al., 2018). For this reason, some authors have advocated for building individual brain models using deep fMRI datasets, where a limited number of individuals get scanned for an extended time, instead of datasets featuring many subjects with limited amounts of data per subject (Naselaris et al., 2021). Second, most fMRI datasets feature only a limited number of stimuli which can be used to train a network. The largest fMRI stimuli sets include Dr Who (approximately 23 hours of video stimuli) (Seeliger et al., 2019) and Natural scenes dataset (10k images) (Allen et al., 2022), which is orders of magnitude smaller than what is currently used in the AI field. For example, the latest version of the recent auditory ANN wav2vec (Baevski et al., 2020) has 317 millions parameters, and was pretrained with over 60k hours of sound data. It thus seems likely that network architectures should feature less parameters when trained for brain alignment than state-of-art networks trained in the field of AI, and the few published studies of brain alignment indeed followed this trend, e.g. (Seeliger et al., 2021; St-Yves et al., 2023). Finally, it is unclear whether training a model directly for brain encoding with a limited set of stimuli will lead to internal representations which do perform well on varied downstream tasks. Although previous works suggest it may be the case (Palazzo et al., 2020; Nishida et al., 2020), those only considered a limited number of downstream tasks: 1 task for Palazzo and 4 for Nishida. These authors also only reported modest gains in performance at best, maybe due to the limited amount of stimuli data used for brain alignment or the limited scope of downstream tasks they considered.

Courtois NeuroMod, HEAR benchmark and model architecture

In this work, we aimed at demonstrating the feasibility of brain alignment of artificial neural networks in the auditory domain. We made several key design choices to address the concerns listed above. First, in line with the approach advocated by Naselaris and

colleagues (Naselaris et al., 2021), we decided to align ANNs at the individual level. We took advantage of the sample collected by Courtois NeuroMod (Boyle et al., 2020), the largest deep fMRI dataset to date, featuring few subjects (N=6) with over 100 hours of fMRI per subject in their 2022 release, and more data yet to be released. The Courtois NeuroMod dataset features a wide range of tasks across multiple domains, and includes several movie watching datasets featuring a complex soundtrack. We focused on the Friends dataset, which is both extensive (we used 36 hours per subject in this study) and varied (it covers multiple seasons of the TV show, with different stimuli at each episode). Second, in terms of **model size and architecture**, instead of training models from scratch, we selected an approach called fine-tuning, where we adjusted the weights of a model which was already pretrained on a large sound dataset. We selected a pretrained model called SoundNet which was both found to be successful for sound processing, brain encoding (Farrugia et al., 2019; Nishida et al., 2020), and also featured a limited number of parameters (less than 3 millions) as well as a simple convolutional architecture, making it amenable to fine-tuning with a limited dataset size. The Friends dataset enabled us both to test the generalisation of brain aligned models to new stimuli in a large controlled distribution (a different season of Friends) and also to replicate the process of brain alignment with six different subjects. Third, in terms of **generalisation abilities on downstream tasks**, we leveraged a recent machine learning competition : the Holistic Evaluation of Auditory Representations (HEAR, Turian et al., 2022). HEAR offers a standardised procedure to test the generalisation of the internal representations of a model on a wide array of downstream tasks. We used the HEAR environment to evaluate brain-aligned models and, as a large number of teams participated in the HEAR competition (Turian et al., 2022), we were able to rigorously compare how our models ranked with a range of state-of-the-art AI approaches.

Study objectives

The specific objectives and hypotheses of our study are as follows:

- Align SoundNet with individual brain data and compare the quality of brain encoding with the baseline, non-brain-aligned model. Our hypothesis was that the alignment procedure would lead to substantial gains in brain encoding for new stimuli drawn from the Friends dataset, and that these gains would be subject-specific, i.e. they would not transfer to other individuals.
- Evaluate how brain alignment impacts downstream tasks. Our hypothesis was that brain alignment would lead to no degradation or even modest improvements in performance across a wide range of tasks.

Taken together, this study establishes the feasibility and some key methodological decisions to align auditory networks with brain data using massive individual fMRI data, and clarifies

how this alignment impacts the performance of networks on downstream tasks.

Materials and Methods

fMRI data

Participants

Six healthy participants (aged 31 to 47 at the time of recruitment in 2018), 3 women (sub-03, sub-04 and sub-06) and 3 men (sub-01, sub-02 and sub-05) were recruited to participate in the Courtois Neuromod Project for at least 5 years. All subjects provided informed consent to participate in this study, which was approved by the ethics review board of the “CIUSS du centre-sud-de-l’île-de-Montréal” (under number CER VN 18-19-22). Three of the participants reported being native francophone speakers (sub-01, sub-02 and sub-04), one as being a native anglophone (sub-06) and two as bilingual native speakers (sub-03 and sub-05). All participants reported the right hand as being their dominant hand and reported being in good general health. Exclusion criteria included visual or auditory impairments that would prevent participants from seeing and/or hearing stimuli in the scanner and major psychiatric or neurological problems. Standard exclusion criteria for MRI and MEG were also applied. Lastly, given that all stimuli and instructions are presented in English, all participants had to report having an advanced comprehension of the English language for inclusion. The above boilerplate text is taken from the cNeuroMod documentation[1] (Boyle et al., 2020), with the express intention that users should copy and paste this text into their manuscripts unchanged. It was released by the Courtois NeuroMod team under the CC0 license.

Friends Dataset

The dataset used in this study is a subset of the 2022-alpha release of the Courtois Neuromod Dataset, called Friends, where the participants have been watching the entirety of seasons 1 to 4 of the TV show Friends. This subset was selected because it provided a rich naturalistic soundtrack, with both a massive quantity of stimuli and a relative homogeneity in the nature of these stimuli, as the main characters of the series remain the same throughout all seasons. Subjects watched each episode cut in two segments (a/b) to allow more flexible scanning and give participants opportunities for breaks. There is a small overlap between the segments to allow participants to catch up with the storyline. The episodes were retransmitted using an Epson Powerlite L615U projector that casted the video through a waveguide onto a blank screen located in the MRI room, visible to the participants via mirror attached to the head coil. Participants wore MRI compatible S15 Sensimetric headphone inserts, providing high-quality acoustic stimulation and substantial attenuation of background noise, and wore custom sound protection gear. More details can be found on the Courtois Neuromod project website[1].

[1]https://docs.cneuromod.ca/en/latest/index.html

Data acquisition

The participants have been scanned using a Siemens Prisma Fit 3 Tesla, equipped with a 2-channel transmit body coil and a 64-channel receive head/neck coil. Functional MRI data was acquired using an accelerated simultaneous multi-slice, gradient echo-planar imaging sequence (Xu et al., 2013) : slice acceleration factor = 4, TR = 1.49 s, TE = 37 ms, flip angle = 52 degrees, voxel size = 2 mm isotropic, 60 slices, acquisition matrix 96x96. In each session, a short acquisition (3 volumes) with reversed phase encoding direction was run to allow retrospective correction of B0 field inhomogeneity-induced distortion.

To minimise head movement, the participants have been provided individual head cases adapted to the shape of their head. Most imaging in the Courtois Neuromod project are composed solely of functional MRI runs. Periodically, an entire session is dedicated to anatomical scans, see details on the Courtois Neuromod project website[1]. Two anatomical sessions were used per subject in this study for fMRIprep anatomical alignment, specifically a T1-weighted MPRAGE 3D sagittal sequence (duration 6:38 min, TR = 2.4 s, TE = 2.2 ms, flip angle = 8 deg, voxel size = 0.8 mm isotropic, R=2 acceleration) and a T2-weighted FSE (SPACE) 3D sagittal sequence (duration 5:57 min, TR = 3.2 s, TE = 563 ms, voxel size = 0.8 mm isotropic, R=2 acceleration).

Preprocessing of the fMRI data

All fMRI data from the 2022-alpha release were preprocessed using the fMRIprep pipeline version 20.2.5 (“long-term support”) (Esteban et al., 2019), see Annex A for details. We used a volume-based spatial normalisation to standard space (MNI152NLin2009cAsym). The anatomical mask derived from the data preprocessing phase was used to identify and select brain voxels. Voxel-level 2D data matrices (TR x voxels) were generated from 4-dimensional fMRI volumes using the NiftiMasker tool from Nilearn (Abraham et al., 2014) and a mask of the bilateral superior temporal gyri middle (middle STG), specifically parcel 153 and 154 of the MIST parcellation (Urchs et al., 2019), resulting in 556 voxels. ROI-level 2D data matrices (TR x ROI) were generated from 4-dimensional fMRI volumes using the NiftiLabelsMasker tool from Nilearn with the MIST parcellation. The MIST parcellation was used as a hard volumetric functional parcellation because of the availability of anatomical labels for each parcel. This functional brain parcellation was also found to have excellent performance in several ML benchmarks on either functional or structural brain imaging (Dadi et al., 2020; Hahn et al., 2022; Mellema et al., 2022). We chose the 210 resolution of the parcellation atlas because parcels were enforced to be spatially contiguous, and separate regions in the left and right hemisphere. Both the middle STG mask used to select the voxels and the parcels from MIST were based on non-linear alignment. The BOLD time series were smoothed spatially with a 5mm Gaussian kernel. A so-called “Minimal” denoising strategy was used to remove confounds without compromising the temporal degrees of freedom, by

regressing out basic motion parameters, mean white matter, and cerebrospinal fluid signals as available in the library load_confounds[2] (equivalent to the default denoising strategy now implemented with load_confounds in Nilearn). This strategy is recommended for data with low levels of motion, as is the case for the Courtois NeuroMod sample (Wang et al., 2023).

Encoding models

Overview

The general approach we adopted to train encoding models of auditory activity relied on transfer learning and fine-tuning of a pretrained deep learning backbone. Audio waveforms were used as inputs to the backbone, generating a set of features as a function of time. Those features were then used to train a small downstream convolutional layer to predict the fMRI signal.

We detail the backbone, fMRI encoding layer, hyperparameters and training procedures below. We used Pytorch and other python packages[3], and trained our models on the Alliance Canada infrastructure on V100 and A100 GPUs.

Deep learning backbone

The network we selected as our backbone is SoundNet, proposed by Aytar and collaborators (Aytar et al., 2016). We selected SoundNet for the following reasons : (1) SoundNet is fully convolutional, as all intermediate representations (i.e. layers) are obtained from 1D convolutions and pooling operators, using directly the audio waveform as input with no additional operations. (2) SoundNet was initially trained on a large dataset of natural videos from the Internet (Flickr), with a high degree of correspondence between the visual and audio content, (3) SoundNet obtained good performances on downstream auditory tasks using transfer learning, as well as good performance as a brain encoding model (Farrugia et al., 2019, Nishida et al., 2020), and (4) SoundNet was trained using a teacher-student training paradigm, in which the student network is fed with raw audio only, and is trained to minimise the discrepancy with probabilities extracted from images using pretrained networks.

At the time of its release in 2016, SoundNet achieved similar performances to the State-of-the-Art (SotA) networks on audio classification benchmarks *Detection and Classification of Acoustic Scenes and Events* DCASE (Mesaros et al., 2017) and *Environmental Sound Classification* ESC-50 (Piczak, 2015; Arandjelovic & Zisserman, 2017). With numerous innovations happening in the AI research field since 2016, as well as the introduction of the dataset AudioSet (Gemmeke et al., 2017), it has since been

[2] https://github.com/SIMEXP/load_confounds
[3] https://github.com/brain-bzh/cNeuromod_encoding_2020/tree/refactoring_V2

surpassed by other networks. However, the CNN architecture dominated the leaderboard of many benchmarks up until 2021 for the audio classification task (Wang et al., 2021; Verbitskiy et al., 2022; Gong et al., 2021), and is still quite relevant with the current considerations of the field to find efficient architectures with fewer parameters and reduced energy consumption (Schmid, Koutini, & Widmer, 2023).

While the naturalistic fMRI dataset we used to fine-tune the network is the biggest up-to-date in the fMRI field, the size of the training dataset is still far below what has been offered by larger audio datasets such as AudioSet or VGGSound (Gemmeke et al., 2017; Chen et al., 2020), often used to train the most recent SotA networks. For this reason, we consider that a smaller, simple network was not only necessary in the context of this study, but beneficial to isolate the impact of brain alignment on network performance. As such, SoundNet provides a generic convolutional network to learn from, with its representations encoding audio features of varying durations, and increasing abstraction in deeper layers.

SoundNet's architecture is a series of convolutional blocks that always include the following steps :

- a 1D Convolutional layer
- a 1D Batch normalisation realised on the output of the convolutional layer
- a rectified linear unit function element-wise ReLU.

In some of the blocks, a 1D max pooling is also applied to the output of the preceding steps (see Table 1 for more details).

fMRI encoding layer

		Initial training input size Audio Input length * TR length (1,49s) * 22050 Hz				***For Sub-03 :*** Audio Input length = 70 TR
Layers		Number of out Channels	Kernel size	Stride	Padding	Number of parameters
SoundNet						
Conv1	Conv1D	16	64	2	32	1040
	BatchNorm1D	16				32
	ReLU	16				
	MaxPool1D	16	8	8		
Conv2	Conv1D	32	32	2	16	16416
	BatchNorm1D	32				64
	ReLU	32				
	MaxPool1D	32	8	8		
Conv3	Conv1D	64	16	2	8	32,832
	BatchNorm1D	64				128
	ReLU	64				
Conv4	Conv1D	128	8	2	4	65,664
	BatchNorm1D	128				256
	ReLU	128				
Conv5	Conv1D	256	4	2	2	131,328
	BatchNorm1D	256				512
	ReLU	256				
	MaxPool1D	256	4	4		
Conv6	Conv1D	141	4	2	2	524,800
	BatchNorm1D	512				1,024
	ReLU	512				
Conv7	Conv1D	1024	4	2	2	2,098,176
	BatchNorm1D	1024				2,048
	ReLU	1024				
Encoding layer Whole-brain						
	Conv1D	210	*Kernel size*	1	*Kernel size -1*	1,075,410
Encoding layer STG						
	Conv1D	556	*Kernel size*	1	*Kernel size -1*	2,847,276

Table 1. Architecture of the SoundNet network with the two different encoding layers. The values for the number of parameters shown on the right side of the table have been estimated for

Sub-03 encoding models, with a selected duration of 70 TRs and a kernel size of 5 for both encoding layers. Values vary slightly for each subject (see Table 2 for the different values used for each subject).

We implemented the SoundNet architecture followed by the fMRI encoding layer as a fully end-to-end trainable network (see Table 1), by adapting an open-source pytorch implementation of SoundNet[4]. The brain encoding layer (as previous soundnet layers) includes temporal 1D convolutional kernels learnt for each pair of inputs and outputs, and the output of each feature map after convolution are summed up to predict brain data for a particular brain parcel/voxel. The output of layer Conv7 is used as input for brain encoding, as this layer has the same temporal resolution as the fMRI signal (0.67 Hz). Formally, the brain encoding layer uses the following model:

$$y_i = \sum_{k=0}^{N-1} h_{i,k} \star x_k$$

where:

- y_i is a window of brain activity associated with parcel/voxel i,
- $N = 2048$ is the number of features in layer 7 (and k denoting a particular feature),
- $h_{i,k}$ is a convolution kernel, analogous to a hemodynamic function, with varying size and parameters trainable specifically for each pair (i, k),
- x_k is a window of activity for the k-th feature of layer 7 of SoundNet, and
- $\star$ is the valid 1D cross-correlation operator (including padding with a size relative to the kernel size, varying between 6 to 9 seconds, or 4 to 6 TR).

Note that we explored multiple lengths (from 20 s up to 130 s) for the time window which we treated as hyper-parameters for optimization (see next section).

Also note that, with a kernel size of 1, this model would be equivalent to a traditional mass-univariate regression of SoundNet features onto brain activity, with no delay between sound waves and brain activity. The proposed model extends this zero-lag regression to incorporate a model of hemodynamic response, with different response functions being trained for each pair of SoundNet feature and brain region.

Targets for brain alignment

The encoding layer was trained using two different brain targets, depending on what type of fMRI processed data the network learned to predict, thus yielding two different models:

- **STG model:** A model to predict fMRI signal from each 556 voxels located in the STG middle mask, at every TR, resulting in a prediction matrix's size of 556 voxels

[4] https://github.com/smallflyingpig/SoundNet_Pytorch

prediction by the selected number of TR (see Table 2 for the exact TRs number for every subject). We refer to this model as the *STG model* in the result section. This model serves as replicatory model, to verify how well SoundNet predict auditory fMRI activity in our settings, and compare its prediction performance to the current literature in auditory encoding models (Kell et al., 2018, Nishida et al., 2020, Caucheteux et al., 2023)

- **Whole-brain model:** A model to predict the average fMRI signal for all voxels of a parcel at every TR, resulting in a prediction matrix's size of 210 parcels (also designed as ROI) prediction by the selected number of TR. We refer to this model as the *Whole-brain model* in the result section. The intention for this model is threefold (1) to verify which brain regions can be predicted by the model using audio as an input, (2) to check which ROIs are impacted by fine-tuning and (3) to test whether individual variability has an impact on prediction performance and fine-tuning.

Data used for	**Training**	**Validation**	**Testing**
Friends	Seasons 1, 2, 3 : 73 episodes		Season 4 : 24 episodes
percentage	75%	25%	-
Numbers of episodes	54,5	18,5	18,5
Total duration (hours)	21,37	7,25	6,98

Table 2. Repartition of the fMRI dataset Friends between model training, validation and testing.

Fixed parameters for model training

To train this architecture, we used AdamW (Loshchilov & Hutter, 2019) as an optimizer for L2 regularisation with weight decay, and we applied a learning rate scheduler that reduces the initial learning rate if no progress is achieved by the optimizer. The weight decay means that the brain encoding layer acts analogously to a Ridge regression, in effect regularising the regression parameters through shrinkage. MSE loss is used to minimise the difference between the predicted and actual fMRI signal.

For training, we used the fMRI data from subjects watching the first three seasons of Friends. For each subject, we used 75 percent of the dataset for training, corresponding to 21 hours of training dataset. The remaining 25 percent was used for validation, corresponding approximately to 7h of dataset, and we use the same amount with Season 4 that was left out to be used only for testing (see Table 2).

To evaluate the accuracy of the model's prediction, we computed the coefficient of determination r^2 between the prediction and the corresponding fMRI time series for each region or voxel, for the entirety of the selected dataset (training or testing).

Hyper-parameters exploration and baseline model

					Early Stopping	
Sub	Audio input length (tr)	Initial learning rate	Kernel size	Weight decay	Patience (epochs)	Delta
03	1	1.10^{-2}	1	1.10^{-2}	10	**0**
	10	1.10^{-3}	3	$\mathbf{1.10^{-3}}$	**15**	0,1
	30	$\mathbf{1.10^{-4}}$	**5**	1.10^{-4}	20	0,5
	70		9			
04	60	1.10^{-3}	4			
	70	1.10^{-4}	5			
	80	$\mathbf{1.10^{-5}}$	**6**	$\mathbf{1.10^{-3}}$	**15**	**0**
06	60	1.10^{-3}	4			
	70	1.10^{-4}	5			
	80	$\mathbf{1.10^{-5}}$	**6**	$\mathbf{1.10^{-3}}$	**15**	**0**
01	60	1.10^{-4}	5			
	70	$\mathbf{1.10^{-5}}$	6			
	80	1.10^{-6}	**7**	$\mathbf{1.10^{-3}}$	**15**	**0**
02	60	1.10^{-4}	5			
	70	$\mathbf{1.10^{-5}}$	6			
	80	1.10^{-6}	**7**	$\mathbf{1.10^{-3}}$	**15**	**0**
05	**60**	1.10^{-4}	5			
	70	1.10^{-5}	6			
	80	$\mathbf{1.10^{-6}}$	**7**	$\mathbf{1.10^{-3}}$	**15**	**0**

Table 3. Hyperparameters values explored for subjects' baseline model. The values selected for fine-tuning the models are marked in bold. The subject order has been constrained by the availability of all four season fMRI data for each subject at the time of the analysis, as this study was taking place in parallel to the data collection. For the kernel size and the learning rate, the range of values shifted

by 1 unit between sub-06 and sub-01, as the optimal values for sub-4 and sub-06 appeared to be the maximal value of the explored window, and we wanted to ensure optimal hyperparameters for each subject.

The goal of this study is to compare performance of an auditory AI network and of the same network but brain-aligned. We decided to realise the hyperparameter grid-search on the original trained SoundNet, with no fine-tuning, which we will consider as a fixed backbone, also referred to as the Baseline model. Through this grid-search, we were looking for an optimal set of hyperparameters to ensure SoundNet prediction performance as an encoding model as well as accounting for individual variability in the fMRI dataset. By going through these optimization steps, we have a better estimation of how much fine-tuning with brain representation impacts a network, for both brain encoding and network performance in classic AI tasks. The selected parameters to be tested only affect the training of the encoding layer at the end of the network.

We optimised different hyperparameters and criteria that could impact the final results (Table 3) :

- the duration of the audio waveforms given as input in each training iteration,
- the value of the learning rate at the beginning of the training,
- the size of the kernels in the encoding layer,
- the initial weight decay,
- the minimal change value considered for early stopping (referred to as “delta”),
- the number of epochs where such delta change is not present before stopping the training (“patience”).

As we are working on individual datasets, we wanted to explore the hyperparameters space for each subject's baseline models. However, doing so would be particularly time-consuming and highly costly in terms of computational resources. In order to constrain the hyperparameters grid search, we decided to explore all parameters only with the baseline model trained on sub-03 fMRI dataset. sub-03 has been chosen amongst others subjects as it was the subject showing the highest Signal-to-Noise Ratio (SNR) in their time series. We trained the baseline model (fixed-weights SoundNet + encoding layer) to predict fMRI activity from sub-03, with 1296 different configurations of hyperparameters (see Table 3 for the selected values explored for each criteria). This step has been done twice, for sub-03 STG model and Whole-brain model, but as results were equivalent between both, we decided to only keep results from the Whole-brain model. To distinguish between all configurations of hyperparameters, we ranked trained models by their configuration and their prediction performance on the validation set: we used the r^2 score of the best predicted parcel as our

measure or performance. We computed a correlation matrix between ROI predictions from the best 100 configurations, to determine if some of the models trained on these configurations shared a similar parcel prediction pattern. We observed 2 to 3 clusters of configurations, so in order to better define these clusters, we used an agglomerative hierarchical clustering, with a linkage function computing the Euclidean distance between centroid of clusters, and divided the output in 3 clusters (UPGMC algorithm, as implemented by the Scipy library[5]). For each cluster, we computed the corresponding predicted brain map by averaging the prediction output of all configurations in the cluster. We then compared all 3 clusters, and chose the cluster with the highest maximal r^2 score (best predicted parcel amongst 210) and highest mean r^2 score (mean of all 210 r^2 scores). When looking at the values for each hyperparameter, we determined which value was prevalent in the best cluster, by computing the statistical mode on the values selected for all configurations present in this cluster (see Supplementary Result 1).

After deciding which parameters and values had the most impact on the capacities of the baseline model to predict fMRI signals from sub-03 dataset, we switched to other subjects, this time exploring only the most impactful hyperparameters on sub-03 and a limited set of values around the optimal value found for sub-03 (see Table 3). As a result, we explored only 27 configurations for each of the remaining subjects. We defined the best hyperparameters values by computing the mode amongst the 10 most performing configurations. We did not use cluster analysis for other subjects, as we only had 27 configurations. As the data collection was done in parallel to the hyperparameter grid-search, we started this process with data from a few subjects (sub-03, sub-04 and sub-06). After selecting the hyperparameters configuration for both sub-06 and sub-04, we decided to adjust the window of tested values of two hyperparameters for the remaining subjects, as their optimal values seemed to always be the highest value for both sub-06 and sub-04 (see Table 3 for more details).

Fine-tuning the model

While only the encoding layer has been trained in the baseline model, the brain-aligned models have part of the original SoundNet's parameters retrained to adjust prediction on individual fMRI data. As our architecture can be trained as an end-to-end network, we decided to test different levels of fine-tuning, from training only the last original layer of SoundNet (Conv7), to training the entirety of the network (Conv1). As such, we obtained 7 fine-tuned models both for Whole-brain and middle STG: Conv1, Conv2, Conv3, Conv4, Conv5, Conv6 and Conv7, each referring to the depth of the network that has been trained.

[5]https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html#scipy.cluster.hierarchy.linkage

Amongst these 7 models, we selected brain-aligned models that have the best ratio between brain encoding performance and training efficiency. We found that models where SoundNet have been fine-tuned up until Conv4 (referred as Conv4 models) achieve the best trade-off (see Supplementary Result 3, figure S2).

Models Comparison and statistical analysis for brain encoding

In order to evaluate the encoding performances of the baseline and fine-tuned models, we tried to predict fMRI activity with a null model, using the same architecture as the other models, but with randomly initialised weights. We used a Wilcoxon test (with a threshold of 0.05) to determine if the difference of the r^2 value of a region / voxel between the null model and the baseline or fine-tuned model was significant across all test runs. As we repeated the same test for 210 regions or 556 voxels, we corrected the p-values obtained through the Wilcoxon test with a False Discovery Rate (FDR), using the Benjamini-Yekutieli procedure (Benjamini & Yekutieli, 2001), to take in account possible dependencies between tests derived at different regions / voxels. Only significant regions with a false-disovery rate q inferior to 0.05 were considered as significant. We repeated the same procedure to determine if the difference of r^2 scores between baseline and fine-tuned models were significant, to evaluate if fine-tuning SoundNet on brain representations had an impact on SoundNet performances.

Evaluating the models on HEAR

To evaluate how fine-tuning impacted SoundNet performances, we tested every brain-aligned and baseline model on the Holistic Evaluation of Audio Representation (HEAR) benchmark (Turian et al., 2022). HEAR was proposed as an AI auditory competition for NeurIPS 2021, and gave the possibility to multiple research teams to test their network architectures and models. This benchmark has been made to evaluate how audio representations from a model are able to generalise over a series of 19 diverse audio tasks, including ESC50 and DCASE 2016, ranging from speech detection to music classification.

As some of the tasks required different inputs, the authors provided an API[6] to adapt existing AI models, in order to be tested with their evaluation code. A wide range of models have been evaluated with this API, resulting in a public ranking of auditory AI models in terms of transfer learning capability at the time of the competition (2022). We extracted the representation of the Conv7 layer to use as scene embeddings, and calculated timestamp embeddings (i.e. a sequence embeddings associated with their occuring times) using the Conv5 layer. For each task of the HEAR benchmark, we quantified the change in ranking for the SoundNet model before vs after fine-tuning with brain data, for each subject separately,

[6] https://hearbenchmark.com/hear-api.html

and for each type of target (Full Brain vs STG). We applied a Wilcoxon test to determine an overall gain (or loss) in ranking across all the 19 tasks available in HEAR for each configuration separately.

Results

Baseline brain encoding using pretrained SoundNet

SoundNet successfully encodes brain activity in the auditory cortex

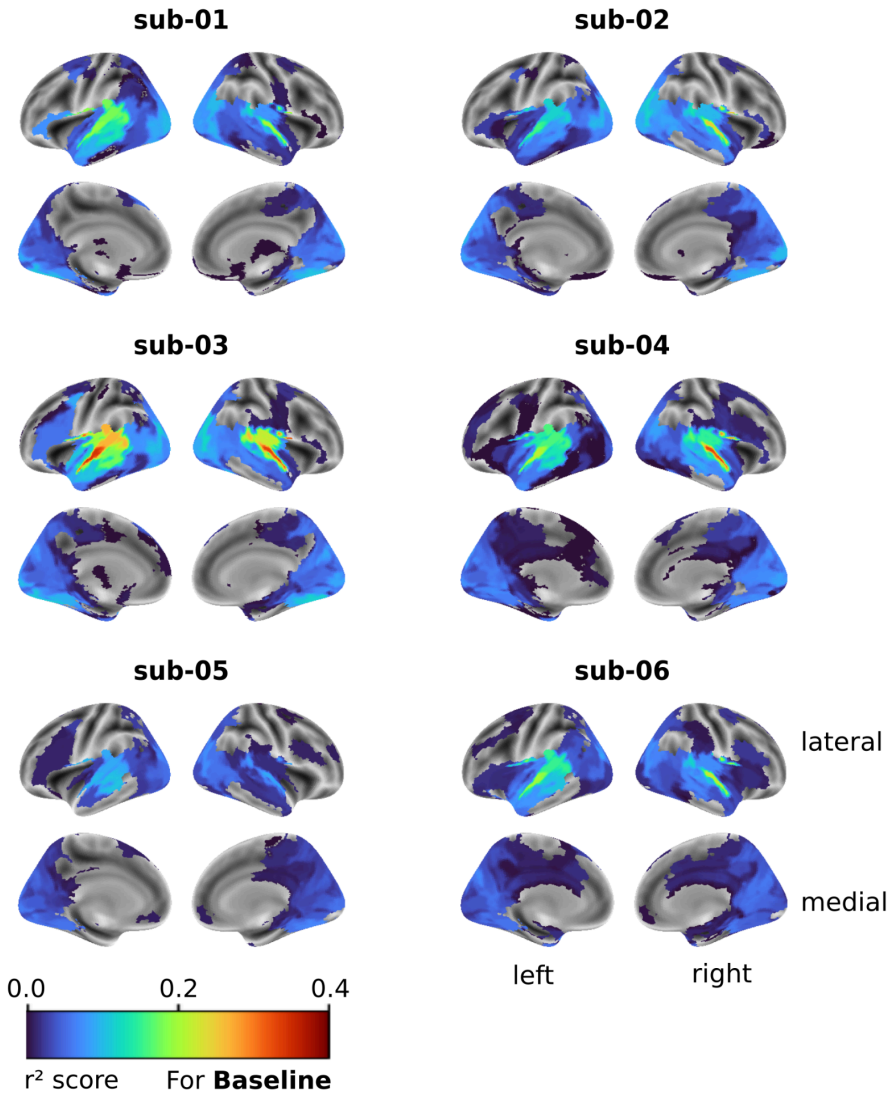

Figure 1. Full brain encoding using SoundNet. Surface maps of each subject, showing the r^2 value for all ROIs from the MIST ROI parcellation. Only parcels with r^2 values significantly higher than those of a null model initialised with random weights are shown (Wilcoxon test, FDR $q < 0.05$). Regions with highest r^2 scores are the STG bilaterally, yet significant brain encoding is achieved throughout most of the cortex, with relatively high values found in the visual cortex as well.

We first tested the ability of our baseline model, SoundNet, to predict fMRI signals in different brain regions. It performed well, especially in the STG. Figure 1 shows almost all subjects had higher r^2 scores in the middle STG (STGm) than other regions ($q < 0.05$ for all subjects), except sub-05 whose best predicted parcel was the Middle Temporal Gyrus (MTG)

superior. The posterior STG (STGp) also consistently ranked second in terms of prediction accuracy (see Supplementary Result 2, Figure S1).

SoundNet also accurately predicted other auditory regions like the MTG and Heschl's gyrus in most subjects (Figure 1). This result supports our hypothesis that our baseline model can encode auditory processing from natural stimuli like movies, with some notable variations in performance between subjects, e.g. substantially higher performance was achieved in STG for sub-03 and sub-04.

SoundNet also encodes brain activity in the visual cortex and other regions

Apart from the auditory cortex, brain activity in other regions was also predicted by the models; for most subjects we observed ROIs in the visual cortex such as the Lateral Visual Network DorsoPosterior (LVISnet_DP) and the Ventral Visual Network (VVISnet), respectively scoring as high as 0.12 and 0.11 (max scores in sub-03). These ROIs proved to be the best predicted ROIs after the STG and the MTG in Sub-01, Sub-02, Sub-05 and Sub-06, revealing that our baseline models were also able to encode aspects of the processing of an audio stimulus outside of the auditory cortex.

SoundNet encodes high resolution brain activity in the Superior Temporal Gyrus

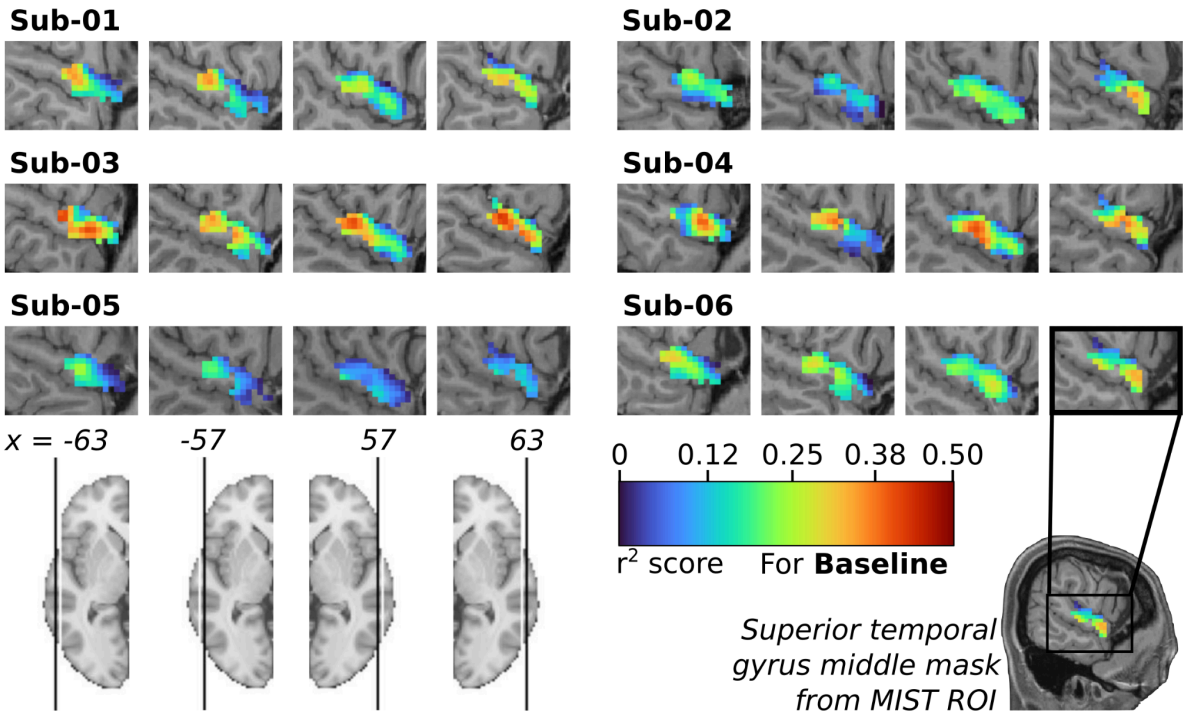

Figure 2. STG encoding using SoundNet. Mapping of the r² scores from 556 voxels inside the cerebral region defined as the Middle STG by the parcellation MIST ROI, computed by the individual Baseline model. To have a better representation of the STG, 4 slices have been selected in each subject, 2 from the left hemisphere (-63 and -57) and 2 from the right hemisphere (63 and 57). Only

voxels with r^2 values significantly higher than those of a null model initialised with random weights are shown (Wilcoxon test, FDR $q < 0.05$). Individual anatomical T1 have been used as background.

We found SoundNet could predict fMRI signals from voxels in the middle STG for all subjects. Most voxels' fMRI activity was accurately predicted, with r^2 scores significantly different from the null model (514 to 555 significant voxels out of 556). Subjects 03 and 04 showed the best performance (average max r^2 of 0.45), while subject 05 performed worse (average max r^2 of 0.27). These results are consistent with the current literature regarding encoding activity from the auditory cortex and SoundNet ability to encode brain activity in the auditory cortex (Caucheteux et al., 2023, Nishida et al, 2020). We saw individual differences in prediction topography: for subjects 01 and 03, efficiency followed an antero-posterior pattern, while other subjects had specific cortical areas where certain voxels were better predicted. These results confirm our model can predict fMRI activity linked to auditory processing at a high spatial resolution. It also shows similar individual differences as observed in full brain encoding.

Fine-tuning SoundNet with individual brain encoding

Individual models do not benefit of the fine-tuning the same way

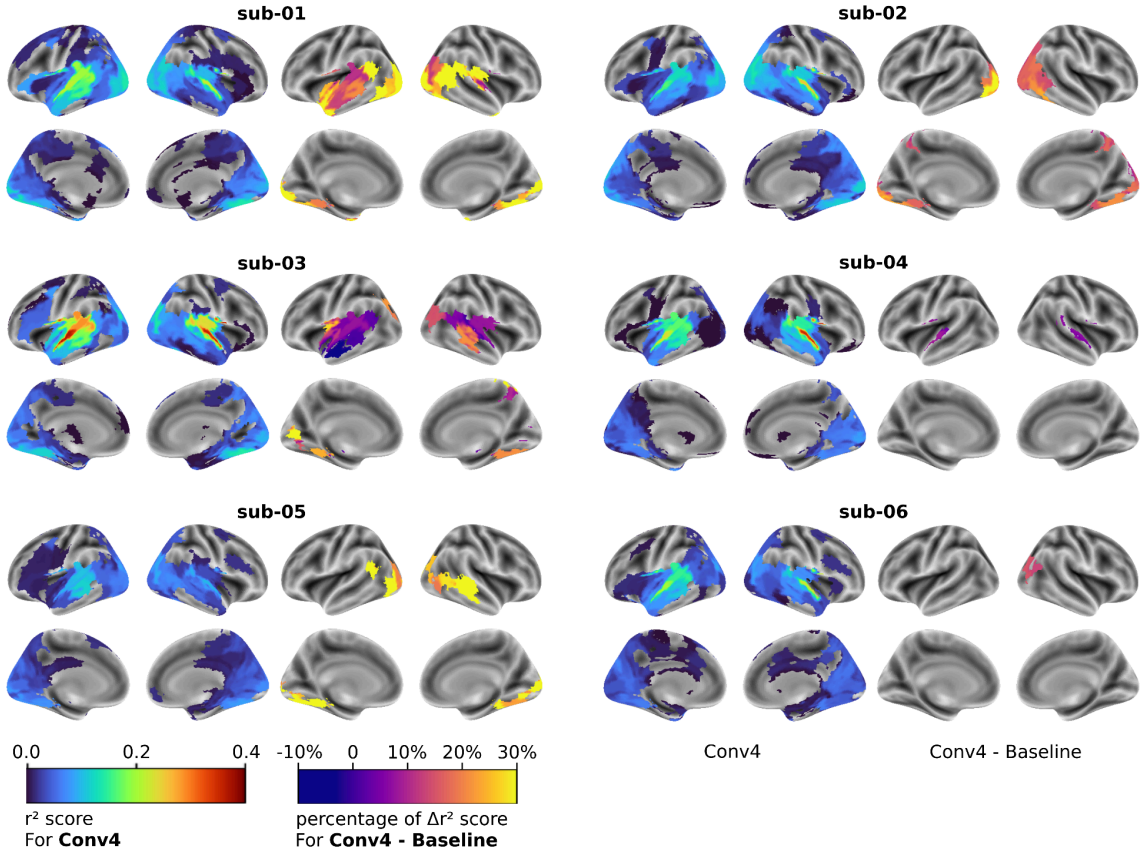

Figure 3. Individual impact of Conv4 fine-tuning on the full brain encoding. For each subject, on the left side : Surface maps of the r^2 scores computed with each individual Conv4 model, for the 210 ROIs of the MIST ROI parcellation. Coloured ROIs have a r^2 score significantly greater than the null model (Wilcoxon test, FDR q <0.05). On the right side : surface maps of the percentage of difference in r^2 scores in each ROI between Conv4 and Baseline models. Only ROIs where Conv4 model have a r^2 score greater than +/- 0.05 and significantly greater or lesser than the baseline model are displayed (Wilcoxon test, FDR q <0.05)

We next examined the fine-tuning impact on the brain-aligned models compared to the baseline models. After fine-tuning, the top-predicted parcels for Conv4 models were the same as those at baseline (Figure 3, left side of each subject panel). For most subjects, STGm and STGp remained the highest-scoring ROIs, with the exception of subject 02 model: while the right STGm is still best predicted, some visual cortices were better encoded than STG regions.

For the number of ROIs predicted by each model, we observed substantial individual variability. Subjects 01, 02, and 05 benefited from fine-tuning with fMRI signals in multiple ROIs, leading to more parcels with significantly different r^2 scores from the null model, compared to results obtained with the baseline models. In contrast, models fine-tuned with subjects 03, 04 and 06's data showed fewer significant ROIs, compared to baseline.

We looked at which brain ROIs had the most improvement using Conv4 fine-tuned models. We tested each ROI's r^2 score for significant difference over the baseline and examined both the r^2 scores difference and the corresponding percentage of difference (see Figure 3, right side for a brain map of the percentage of r^2 score difference for each ROI, between the Conv4 and Baseline model for each subject):

The fine-tuning's main improvements were not always in the auditory cortex: while for subjects 01, 04 and 05, the highest gain in r^2 score was in the right STG or MTG (between +0.02 and 0.04 r^2 score), for the remaining subjects it was located in the ventral, lateral or posterior visual network. In general, we observed important r^2 score improvements in the visual cortex in most subjects, except for sub-04 model where the prediction has been improved only in the auditory cortex, and the model's predicting capacity in the frontal and parietal cortex was even degraded by fine-tuning.

For most subjects, ROIs with the highest improvement in r^2 score gained between 0.01 and 0.04, with a relative gain from their original value of 15% to 30%, depending on the ROI and the subject. However, Sub-05 brain-aligned model, whose baseline model had the worse r^2

scores amongst all subjects, showed the highest relative gain in the MTG posterior (+ 0,03 r^2 score, corresponding to a gain of 167% of the original value). In general, ROIs with low r^2 scores (between 0.05 and 0.15) showed higher relative improvement than ROI with high r^2 score.

Overall, fine-tuning improved the quality of brain encoding overall, with substantial variations across subjects both in terms of the magnitude and location of improvements.

Fine-tuning at the voxel level also leads to substantial improvements in brain encoding

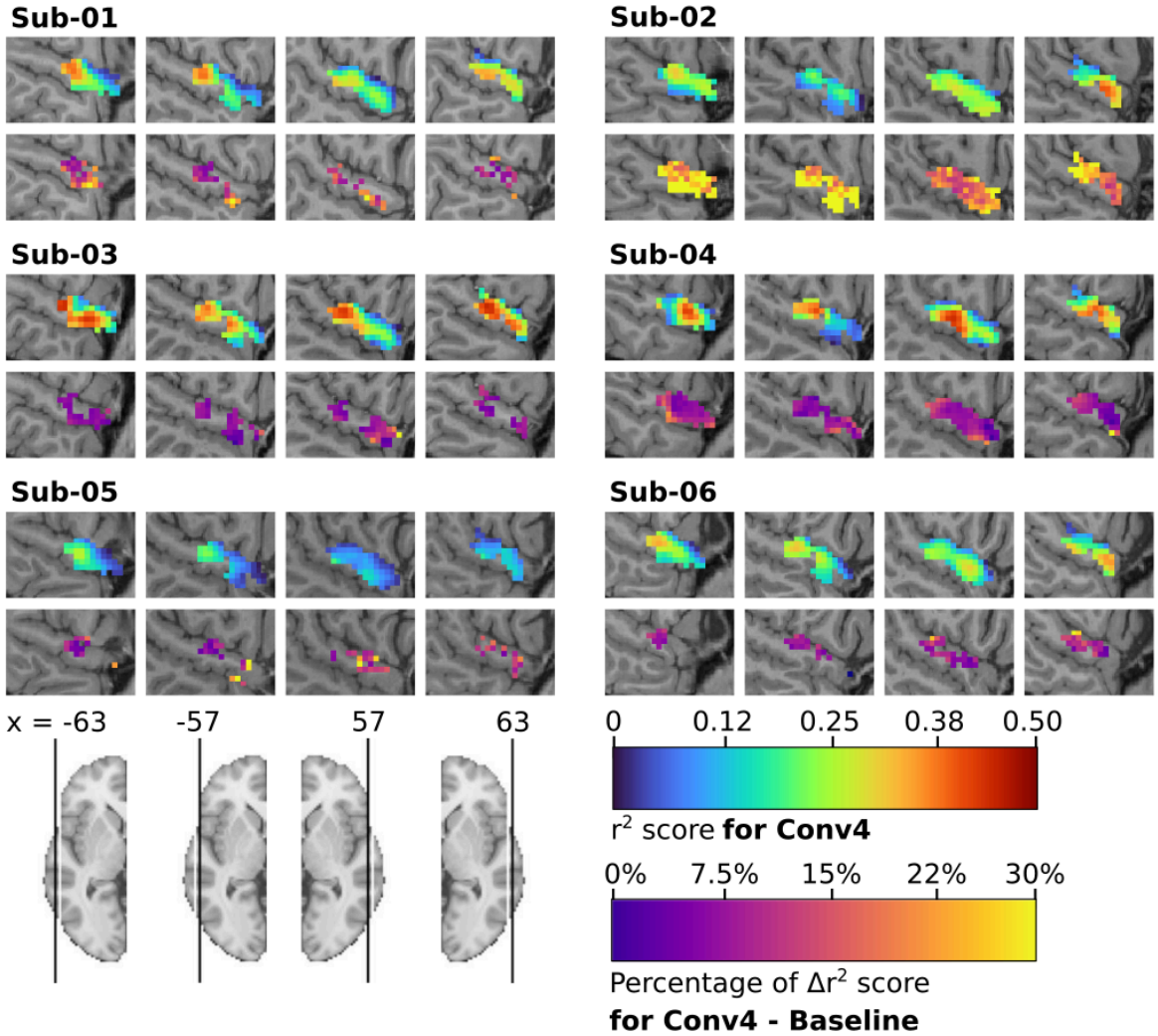

Figure 4. STG encoding using SoundNet. For each subject, on the top : mapping of the r^2 scores from 556 voxels inside the cerebral region defined as the Middle STG by the parcellation MIST ROI, computed by the individual Conv4 model. Only voxels with r^2 values significantly higher than those of a null model initialised with random weights are shown (Wilcoxon test, FDR $q < 0.05$). For each subject, on the bottom : mapping of the difference of r^2 scores between the Conv4 model and the baseline model. Only voxels from the Conv4 model with r^2 values greater than +/- 0.05 and significantly greater or lesser than those of the baseline model are shown (Wilcoxon test, FDR $q < 0.05$). Individual anatomical T1 have been used as background.

We next wanted to see if fine-tuning also affected voxel-level fMRI signal predictions. We calculated the r² score difference between the baseline and Conv4 for each voxel in the STGm ROI, and only mapped those with significant differences (Figure 4). Voxels that were well predicted by the baseline models also had the most significant r² score increases for all subjects, although voxels with lower prediction accuracy were also impacted by fine-tuning. However, the relative gain for most voxels and subjects was between 6 to 8%, lower than relative gains found in the brain-aligned WB models. The fine-tuning improved subject 05's model less than others, as indicated by a smaller general r² score increase. The notable exception is subject 02 STG model, where many voxels demonstrated a relative gain superior to 30%, with most of these voxels originally having a lower r² score in the baseline model. These findings align with Whole-brain model results. Overall, we found that fine-tuning improved brain encoding at the voxel level, with marked variations across subjects, and some departure from the impact of fine-tuning at the level of the full brain.

Fine-tuning improves SoundNet ranking in diverse AI auditory benchmarks

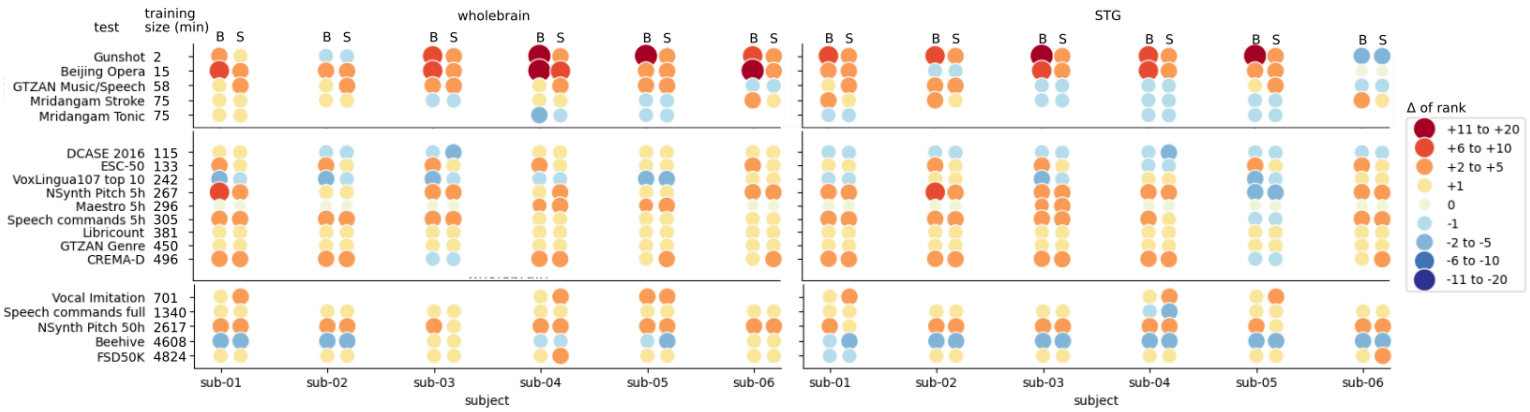

Figure 5. Rank variation between Conv4 and baseline models on all tasks from the HEAR benchmark. Each individual Conv4 model (both Whole-brain and middle STG models) has been used to resolve the 19 tasks from the HEAR Benchmark, ordered by the size of training dataset available through the benchmark. We compared Conv4 models performance to SoundNet against 8 small models, with less than 12 millions parameters (S columns, on the right side for each subject), and 21 big models (B columns, on the left side for each subject). For each task and each subject, the change of rank between the baseline model and the Conv4 model is symbolised by a coloured circle. When the change of rank is equal to +1, Conv4 model is performing better than SoundNet at the task, but doesn't outperform other models.

We aimed to evaluate the impact of brain alignment on the performance of SoundNet with downstream tasks, using the HEAR benchmark. For each task in the benchmark, we

compared Whole-brain and STG fine-tuned models for each subject against SoundNet, and ranked both brain-aligned models and Soundnet amongst the 29 other models tested with this benchmark. In order to compare brain-aligned models performance with models of similar size (around 3 millions parameters), we divided the 29 models in two evaluation groups, depending of the number of parameters: a first group of 8 small models having less than 12 millions parameters, and a group with the remaining 21 models, ranging from 22 millions to 1339 millions parameters for the bigger models. We also divided tasks depending on the dataset size used for training, to evaluate how brain-aligned models performed when having access to a limited range of data to generalise to a new task or to a bigger training dataset.

After evaluating SoundNet in each task using the HEAR API, we found that SoundNet didn't perform well in most tasks, being part of the least performing models multiple times. Brain-aligned models (all subjects, both WB and STG) performed significantly better than SoundNet ($p < 0.05$) in 11 tasks out of 19, and performed worst in 1 task (Beehive) (see Supplementary Result 4, figure S3 and Table S4).

Using Whole-brain fine-tuning, all 6 individual brain-aligned models displayed a significant gain in rank in the benchmark (over SoundNet and other models) ($p < 0.05$) (left panel of the figure 5). When comparing how brain-aligned models ranked amongst big models and small models (respectively *B* column and *S* column, for each individual model in figure 7), results were similar: all individual models had a significant increase in gain rank amongst small models, and 5 individual models displayed a similar increase amongst big models, with the exception of Whole-brain Sub-2 model, where its increase was still close to be significantly different ($p = 0.057$).
Results with individual middle STG models were more heterogeneous, with only sub-01, sub-02 and sub-03 Conv4 models ranking significantly higher than SoundNet, amongst small models and big models (right panel of Figure 5)

If we consider improvement only over other models than SoundNet, we observed a significant gain in rank only for a few subjects for both Whole-brain fine-tuning and STG fine-tuning. This result is due to the fact that in multiple tasks, Brain-aligned models only surpass SoundNet. However, in some case, the fine-tuning led to important gain in rank, as specific tasks seem to benefit much more from the fine-tuning than others. Examples of tasks benefitting a lot from fine-tuning include Gunshot Triangulation, Beijing Opera and the NSynth Pitch (5h and 50h), while DCase 2016 was not improved. Performance in these tasks also vary between individual models, and between middle STG and Whole Brain models. Brain alignment had the biggest impact with tasks involving small training datasets :

Brain-aligned models surpassed up to 18 others models (both small and big models included), while they only had 2 minutes to retrain the encoding layer necessary to solve the task. Beijing opera shows similar results, but also has the highest standard deviation in ranking amongst all tasks. Considering that most models tested in this task scored around 0.95 in accuracy, we also consider that the change in ranks observed for Beijing Opera is highly affected by the ranking distribution (see Supplementary Result 4, figure S3).

Overall, aligning SoundNet with brain data led to modest improvements on multiple and varied downstream tasks, with some heterogeneity across the tasks, subjects and methods of fine-tuning (Whole-brain vs STG).

Fine-tuned Conv4 models are subject specific

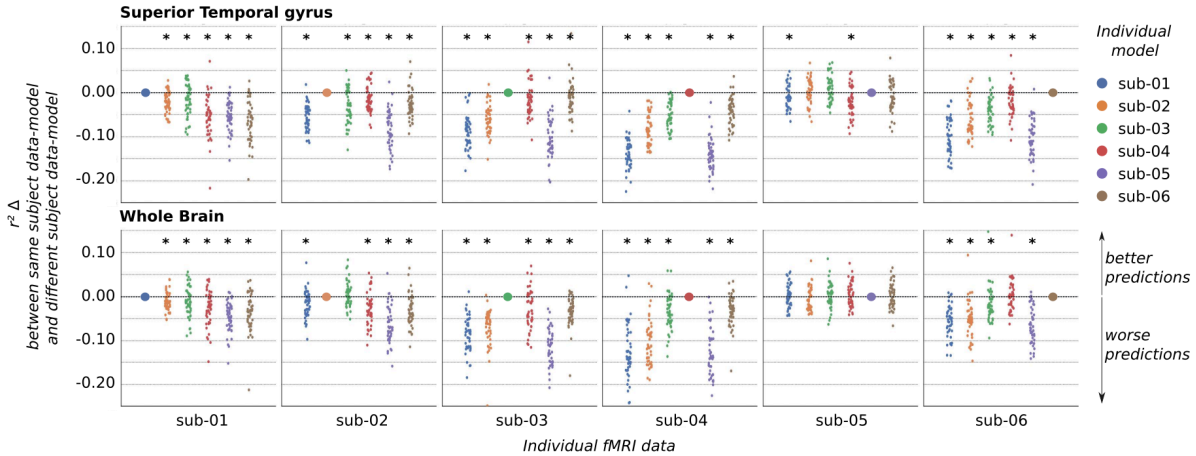

Figure 6. Difference of r² scores computed by an individual Conv4 model, predicting fMRI activity from the same subject that the data used to train it, versus predicting fMRI activity from another subject. The difference is computed for each of the 48 runs of the fourth season of Friends. A Wilcoxon test has been used to determine if the difference was significant between one subject and each of the other 5 subjects.

Finally, we evaluated if the fine-tuned models were subject-specific, by applying models trained on data from one subject to fMRI signals collected on other subjects. We found that the model trained on an individual had the best performance to predict this subject's data, both the Whole-Brain and middle STG models, with the exception of Sub-05 (figure 6). Overall, trained models appeared to exhibit subject-specific features.

Discussion

In this study, we explored the benefits of aligning a pretrained auditory neural network (SoundNet) with individual brain activity, both in terms of generalisation of brain encoding to new types of stimuli, and behavioural performance on a wide range of downstream tasks. Our results confirm substantial improvements in encoding brain activity, with gains extending beyond the auditory cortex, e.g. in the visual cortex. Importantly, brain alignment led to significant enhancements in performance across a broad range of auditory tasks when assessed using transfer learning. Our study also highlighted notable inter-individual variations, both in terms of the impact of brain alignment on brain encoding quality, and in the performance gains for downstream tasks.

Task performance on downstream tasks

We evaluated the performance of our brain-aligned models against SoundNet using the HEAR benchmark, which encompasses a variety of auditory tasks, and found that brain alignment generally benefited performance on downstream tasks. Few studies have employed a downstream task benchmark after brain alignment. Palazzo et al. (2020) reported modest performance gains in vision tasks post-alignment with EEG data. Nishida et al. (2020) reported similar findings with their audiovisual tasks, using fMRI. However, Schwartz et al. (2019) reported no significant change in performance after brain alignment with fMRI or MEG data. Our research differs notably in stimulus nature (a TV show), and the very large volume of fMRI data used for fine-tuning. While our results seem to align with the first two studies in terms of finding mostly moderate improvements in performance, this study is the first to examine a wider range of auditory downstream tasks. The primary goal of the HEAR benchmark is to evaluate the capacity of a network's internal representations to generalise to new tasks with data of a different nature than what has been used to initially train the network. Considering this goal, brain-aligning a pre-trained CNN network led to more generalisable representations, but also identify possibly large gains for downstream tasks with limited training data available: The two tasks that benefited the most are gunshot triangulation (a classification task) and Beijing Opera percussion (an instrument recognition task), which are small scale datasets (Turian et al., 2022) (training data corresponds to approx. 100s and 900s, respectively). However, our results also show improvements on much larger datasets, such as NSynth pitch classification on 5h and 50 hours of data (Engel et al., 2017), as well as modest benefits on a very large and difficult benchmark, FSD50k, a multi-label audio tagging dataset with more than 80 hours of training data (Fonseca et al., 2021). Taken together, the ability of our brain-aligned representations to generalise to small and larger scale datasets suggest both that they are general enough to generalise with few data, and flexible enough to enable gains on larger scale tasks.

While models aligned to Whole-brain data generally outperformed those aligned only to the middle STG, task-specific performance sometimes varied significantly depending on the choice of brain targets. For example, Gunshot triangulation, for which the models need to correctly identify the location of the microphone used to record the sound, showed more substantial improvements for Whole-brain models.For Nsynth Pitch and Speech commands, where models need to classify pitch and speech respectively, both tasks benefited more with middle STG models. This variability was also apparent across individual subjects, as large improvements on certain tasks were not observed systematically for all subjects.

Further research is needed to understand the sources of this variability, and clarify which aspects of the tasks and brain activity are critical to benefit from brain alignment. The CNeuroMod data collection includes a variety of stimuli beyond the friends TV show, and it would be possible to check how different types of stimuli impact generalisation to downstream AI tasks. Additionally, we're interested in investigating whether brain-aligned models lead to human-like similarity judgement patterns (Bakker et al., 2022).

We recognize that the TV show used in the dataset for this study presents some similarities between each season, and may possibly lack in diversity to pretrain an ANN and check for broad generalisation of representations.

Can task-optimized ANN be aligned with brain activity ?

Our findings suggest task-optimised ANNs can successfully be aligned with individual brain activity. While we observed substantial enhancements in brain encoding quality, the extent of these improvements varied across both brain regions and individuals. When models were directly fine-tuned to encode voxel-level STG activity, all participants exhibited modest but significant improvement in the superior temporal gyrus (STG). When models were fine-tuned on the entire brain, the STG remained the best predicted region in the brain after brain-alignment. However, the impact of this process varied across both brain regions and individuals. For most subjects, regions outside the STG such as the visual cortex, experienced improvements comparable or greater to those in the STG.

Different reasons can explain this result: Activation of the visual cortex by auditory stimuli have been observed in different context (Wu et al., 2007; Cate et al., 2009), and as the individual brain-aligned models are highly specific of the individual data used for the training, it is possible that these results directly reflect specificity of individual processing learned with the Brain Alignment. However, it could also be due to the correlation of audio and visual

features in our video stimuli (for instance, the presence of faces and lip movements during speech). It is challenging to draw direct comparisons with previous studies due to the sensitivity of the R2 metric to data acquisition and preprocessing decisions, including smoothing and voxel resolution.

Limitations

A limitation of this study is to focus on a single pretrained network, Soundnet. Considering recent advances in AI auditory models performance (Schmid, Koutini, & Widmer, 2023), it would be important to study other architectures as well in the future, to evaluate if and how brain-alignment impact could differ depending on the architecture used (e.g. Transformers versus convolutional networks), the number of model parameters and the type of data used for pretraining. We also found that SoundNet had a lower score in benchmarks such as ESC-50 and DCase, compared to the scores of its original paper. While we tried to stay as close as the original implementation, multiple reasons could explain this difference of score: it is possible that the original SoundNet paper used different embeddings to evaluate the model, compared to the embedding required by the HEAR benchmark. We also used Python with Pytorch to implement the brain-aligned models and end-to-end training, while originally SoundNet was done in Lua with Tensorflow. The conversion from one library to another could also have an impact.

Another limitation resides in the choice of the testing dataset: while Season 4 of Friends might be considered as an entirely new set of input, the auditory stimuli still displays important similarities with the previous seasons, in particular a very large representation of speech with the same actor voices throughout the four seasons. Further research could test the generalisation of brain encoding to an entire new class of auditory stimuli, which will be possible in particular with the full CNeuroMod dataset.

It should finally be noted that the parcels used from the MIST ROI parcellation were based on non-linear alignment; while the models trained on the Whole-brain fMRI activity best predicted the STG, they also displayed important individual differences. We can’t exclude the possibility that specific ROIs in the auditory cortex and the visual cortex could be slightly misaligned with individual anatomy, which could partially impact the results.

Conclusions

In our study, we developed the first set of auditory deep artificial networks fine-tuned to align with individual participants' brain activity. This was made possible by the Courtois NeuroMod project's massive individual data collection effort. We successfully fine-tuned a pretrained

network called SoundNet to better encode individual participants' brain signals, showing varying degrees of improvement over a model that only adds an encoding layer to predict brain signals. These brain-aligned models also improved in performance a pretrained network, trained without brain data on a diverse set of AI audio tasks, ranging from classifying pitch to determining the number of speakers. The brain-aligned models also demonstrate high potential for tasks with limited dataset available and few-shot learning. These findings open many avenues for future research, ranging from studying inter-individual variations to testing brain alignment for various model architectures, types of training data and types of downstream tasks.

Ethics

All subjects provided informed consent to participate in this study, which was approved by the ethics review board of the “CIUSS du centre-sud-de-l’île-de-Montréal” (under number CER VN 18-19-22).

Data and Code Availability

The fMRI data used to train the model is openly available through registered access at link https://www.cneuromod.ca/access/access/. All the code used to train the models, produce the results and figures are freely accessible on a github repository at https://github.com/brain-bzh/cNeuromod_encoding_2020.

Authors Contributions

Prototyping, development, and fine-tuning of the models have been done by MF and NF, with advice from PB.

API development to adapt models for evaluation with the HEAR benchmark has been done by NF, LT and MLC.

Statistical analysis and visualisation has been done by MF, with advice from NF and PB.

Manuscript has been written by MF and PB, with additional reviewing and contributions on the methods and discussion section by NF. The AI tool ChatGPT4 has been used only to help in improving text architecture and writing style.

Funding and Acknowledgments

The Courtois project on neural modelling was made possible by a generous donation from the Courtois foundation, administered by the Fondation Institut Gériatrie Montréal at CIUSSS du Centre-Sud-de-l’île-de-Montréal and University of Montreal. The Courtois NeuroMod team is based at “Centre de Recherche de l’Institut Universitaire de Gériatrie de Montréal”, with several other institutions involved. See the cNeuroMod documentation for an up-to-date list of contributors (https://docs.cneuromod.ca).

The work was partly supported by a grant from the Brittany region in France “Allocation de Recherche Doctorale” to NF, and a Digital Alliance Canada resource allocation grant to PB to access the Béluga high performance computing infrastructure. PB is supported by a salary award as senior fellow (chercheur boursier senior) of the “Fonds de Recherche du Québec - Santé”.

Competing Interests

The authors declare no competing interests.

References

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., & Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, *8*, 14.

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., ... & Kay, K. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. Nature neuroscience, 25(1), 116-126.

Arandjelovic, R., & Zisserman, A. (2017). Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision* (pp. 609-617).

Aytar, Y., Vondrick, C., & Torralba, A. (2016). Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, *29*.

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

Bakker, M. A., Chadwick, M. J., Sheahan, H. R., Tessler, M. H., Campbell-Gillingham, L., Balaguer, J., McAleese, N., Glaese, A., Aslanides, J., Botvinick, M. M., & Summerfield, C. (2022). *Fine-tuning language models to find agreement among humans with diverse preferences*. arXiv. https://doi.org/10.48550/arXiv.2211.15006

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798–1828.

Benjamini, Y., & Yekutieli, D. (2001). The Control of the False Discovery Rate in Multiple Testing under Dependency. *Annals of Statistics*, *29*(4), 1165–1188.

Boyle, J. A., Pinsard, B., Boukhdhir, A., Belleville, S., Brambatti, S., Chen, J., Cohen-Adad, J., Cyr, A., Fuente, A., Rainville, P., & Bellec, P. (2020, June). *The Courtois project on neuronal modelling - 2020 data release*. Poster 1939 was presented at the 2020 Annual Meeting of the Organization for Human Brain Mapping, held virtually.

Cate, A. D., Herron, T. J., Yund, E. W., Stecker, G. C., Rinne, T., Kang, X., ... & Woods, D. L.

(2009). Auditory attention activates peripheral visual cortex. PloS one, 4(2), e4645.

Caucheteux, C., Gramfort, A., & King, J.-R. (2023). Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behaviour*, 1–12.

Chen, H., Xie, W., Vedaldi, A., & Zisserman, A. (2020, May). Vggsound: A large-scale audio-visual dataset. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 721-725). IEEE.

Conwell, C., Prince, J. S., Alvarez, G. A., & Konkle, T. (2021, October). What can 5.17 billion regression fits tell us about artificial models of the human visual system?. In *SVRHM 2021 Workshop@ NeurIPS*.

Dadi, K., Varoquaux, G., Machlouzarides-Shalit, A., Gorgolewski, K. J., Wassermann, D., Thirion, B., & Mensch, A. (2020). Fine-grain atlases of functional modes for fMRI analysis. *NeuroImage*, *221*, 117126.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Engel, J., Resnick, C., Roberts, A., Dieleman, S., Norouzi, M., Eck, D., & Simonyan, K. (2017, July). Neural audio synthesis of musical notes with wavenet autoencoders. In *International Conference on Machine Learning* (pp. 1068-1077). PMLR.

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., & Gorgolewski, K. J. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), 111–116.

Farrugia, N., Nepveu, V., & Villamil, D. C. A. (2019, October). Estimating encoding models of cortical auditory processing using naturalistic stimuli and transfer learning. In *Real Neurons & Hidden Units: Future directions at the intersection of neuroscience and artificial intelligence@ NeurIPS 2019*.

Fonseca, E., Favory, X., Pons, J., Font, F., & Serra, X. (2021). Fsd50k: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 829-852.

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., ... & Ritter, M. (2017, March). 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). IEEE.

Giordano, B. L., Esposito, M., Valente, G., & Formisano, E. (2023). Intermediate acoustic-to-semantic representations link behavioral and neural responses to natural sounds. *Nature Neuroscience*, 1–9.

Gong, Y., Chung, Y. A., & Glass, J. (2021). Psla: Improving audio tagging with pretraining, sampling, labeling, and aggregation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 3292-3306.

Gordon, E. M., Laumann, T. O., Gilmore, A. W., Newbold, D. J., Greene, D. J., Berg, J. J., Ortega, M., Hoyt-Drazen, C., Gratton, C., Sun, H., Hampton, J. M., Coalson, R. S., Nguyen, A. L., McDermott, K. B., Shimony, J. S., Snyder, A. Z., Schlaggar, B. L., Petersen, S. E., Nelson, S. M., & Dosenbach, N. U. F. (2017). Precision functional mapping of individual human brains. *Neuron*, *95*(4), 791–807.e7.

Gratton, C., Laumann, T. O., Nielsen, A. N., Greene, D. J., Gordon, E. M., Gilmore, A. W., Nelson, S. M., Coalson, R. S., Snyder, A. Z., Schlaggar, B. L., Dosenbach, N. U. F., & Petersen, S. E. (2018). Functional Brain Networks Are Dominated by Stable Group and Individual Factors, Not Cognitive or Daily Variation. *Neuron*, *98*(2), 439–452.e5.

Hahn, S., Owens, M. M., Yuan, D., Juliano, A. C., Potter, A., Garavan, H., & Allgaier, N. (2022). Performance scaling for structural MRI surface parcellations: a machine learning analysis in the ABCD Study. *Cerebral Cortex* , *33*(1), 176–194.

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, *98*(3), 630–644.e16.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*,

25.

Loshchilov, I., & Hutter, F. (2019). *Decoupled Weight Decay Regularization*. arXiv. https://doi.org/10.48550/arXiv.1711.05101

Lu, Z., Wang, Y., & Golomb, J. D. (2024). ReAlnet: Achieving More Human Brain-Like Vision via Human Neural Representational Alignment. *arXiv preprint arXiv:2401.17231*.

Mellema, C. J., Nguyen, K. P., Treacher, A., & Montillo, A. (2022). Reproducible neuroimaging features for diagnosis of autism spectrum disorder with machine learning. *Scientific Reports*, *12*(1), 3057.

Mesaros, A., Heittola, T., Benetos, E., Foster, P., Lagrange, M., Virtanen, T., & Plumbley, M. D. (2017). Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(2), 379-393.

Naselaris, T., Allen, E., & Kay, K. (2021). Extensive sampling for complete models of individual brains. *Current Opinion in Behavioral Sciences*, *40*, 45–51.

Nishida, S., Nakano, Y., Blanc, A., Maeda, N., Kado, M., & Nishimoto, S. (2020, April). Brain-mediated transfer learning of convolutional neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 04, pp. 5281-5288).

Palazzo, S., Spampinato, C., Kavasidis, I., Giordano, D., Schmidt, J., & Shah, M. (2020). *Decoding Brain Representations by Multimodal Learning of Neural Activity and Visual Features*. arXiv. http://arxiv.org/abs/1810.10974

Piczak, K. J. (2015, October). ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia* (pp. 1015-1018).

Schmid, F., Koutini, K., & Widmer, G. (2023, June). Efficient large-scale audio tagging via transformer-to-cnn knowledge distillation. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., Schmidt, K., Yamins, D. L. K., & DiCarlo, J. J. (2020). *Brain-Score: Which Artificial Neural Network for Object Recognition is most*

Brain-Like? bioRxiv. https://doi.org/10.1101/407007

Schwartz, D., Toneva, M., & Wehbe, L. (2019). Inducing brain-relevant bias in natural language processing models. *Advances in neural information processing systems*, *32*.

Seeliger, K., Sommers, R. P., Güçlü, U., Bosch, S. E., & Van Gerven, M. A. J. (2019). A large single-participant fMRI dataset for probing brain responses to naturalistic stimuli in space and time. BioRxiv, 687681.

Seeliger, K., Ambrogioni, L., Güçlütürk, Y., van den Bulk, L. M., Güçlü, U., & van Gerven, M. A. J. (2021). End-to-end neural system identification with neural information flow. *PLoS Computational Biology*, *17*(2), e1008558.

St-Yves, G., Allen, E. J., Wu, Y., Kay, K., & Naselaris, T. (2023). Brain-optimized deep neural network models of human visual areas learn non-hierarchical representations. *Nature Communications*, *14*(1), 3329.

Turian, J., Shier, J., Khan, H. R., Raj, B., Schuller, B. W., Steinmetz, C. J., ... & Bisk, Y. (2022, July). Hear: Holistic evaluation of audio representations. In *NeurIPS 2021 Competitions and Demonstrations Track* (pp. 125-145). PMLR.

Urchs, S., Armoza, J., Moreau, C., Benhajali, Y., St-Aubin, J., Orban, P., & Bellec, P. (2019). MIST: A multi-resolution parcellation of functional brain networks. *MNI Open Research*, *1*, 3.

Verbitskiy, S., Berikov, V., & Vyshegorodtsev, V. (2022). Eranns: Efficient residual audio neural networks for audio pattern recognition. Pattern Recognition Letters, 161, 38-44.

Wang, M., Chen, C., Xie, Y., Chen, H., Liu, Y., & Zhang, P. (2021). Audio-visual scene classification using transfer learning and hybrid fusion strategy. DCASE2021 Challenge, Tech. Rep, Tech. Rep.

Wang, H.-T., Meisler, S. L., Sharmarke, H., Clarke, N., Gensollen, N., Markiewicz, C. J., Paugam, F., Thirion, B., & Bellec, P. (2023). Continuous Evaluation of Denoising Strategies in Resting-State fMRI Connectivity Using fMRIPrep and Nilearn. *bioRxiv : The Preprint Server for Biology*. https://doi.org/10.1101/2023.04.18.537240

Wu, C. T., Weissman, D. H., Roberts, K. C., & Woldorff, M. G. (2007). The neural circuitry

underlying the executive control of auditory spatial attention. Brain research, 1134, 187-198.

Xu, J., Moeller, S., Auerbach, E. J., Strupp, J., Smith, S. M., Feinberg, D. A., ... & Uğurbil, K. (2013). Evaluation of slice accelerations using multiband echo planar imaging at 3 T. *Neuroimage*, *83*, 991-1001.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.