Aligning the activity of artificial and biological neural networks to build personalised models of auditory processing in a massive individual fMRI dataset

Maelle Freteault[1,2,3]*, Loic Tetrel[2,4], Maximilien Le Clei[2], 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 demonstrate here the creation of personalised auditory artificial neural models directly aligned with individual brain activity. 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. We fine-tuned SoundNet, an established deep artificial neural network, to achieve substantial improvement in predicting individual brain activity, including but not limited to the auditory and visual cortices. Performance gains on the HEAR evaluation benchmark, a large collection of downstream AI audio tasks, were modest but consistent, demonstrating that brain alignment leads to more generalizable representations. The fine-tuned models were also subject-specific, as models trained on a specific subject outperformed models trained from other subjects' data. The resulting individual brain models thus provide a new tool for exploring the idiosyncratic representations of auditory signals within the individual human brain.

Keywords

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

Introduction

Overall objective

Auditory neuroscience aims to understand how the brain processes sounds, and how this specific information is morphed into abstract concepts shared across multiple sensory and control streams, such as vision and language. However, accurate modelling of such complex processing remains elusive. Artificial neural networks (ANNs) trained on complex tasks, such as speech recognition using naturalistic sounds, have successfully predicted brain activity captured by functional magnetic resonance imaging (fMRI), thus providing novel insights into brain function. While substantial progress has been made, existing research predominantly aligns ANNs with averaged brain activity across multiple subjects, thereby overlooking individual-specific features of brain organisation. The past decade has seen a push towards deep, individual-focused neuroimaging datasets. These have opened the possibility of directly aligning ANNs with individual brain responses, thus providing a more personalised understanding of auditory processing. This approach not only enables the characterization of individual-specific features of brain organisation but may also enhance the generalisation of learned representations to downstream tasks. In the present study, we aim to align a pretrained auditory ANN with individual brain data and study the subsequent impact on auditory task performance, using a rich, deep fMRI dataset from the Courtois NeuroMod project (Boyle et al., 2020).

ANNs for brain encoding in audition

ANNs trained with deep learning for cognitive neuroscience were initially in the field of vision (Schrimpf et al., 2020). Convolutional neural networks (CNN) 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 (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). However, these studies typically used a single pretrained ANN to predict the brain activity of multiple subjects, reflecting an underlying assumption of shared features across all subjects' brain activity. This approach, while practical, overlooks substantial evidence of individual differences in brain representations (Gordon et al., 2017; Gratton et al., 2018; Naselaris et al., 2021), indicating a need for individual brain models.

Individual brain models

Cognitive neuroscience has seen an uptick in the use of deep, individualised datasets, such as the myConnectome project, wherein an individual underwent rigorous scanning for 18 months (Poldrack et al., 2015), and the Natural Scenes Dataset (Naselaris et al., 2021), featuring exhaustive imaging of up to 40 hours per subject (N=10). The NSD in particular was established with a unique emphasis on tailoring artificial neural networks (ANNs) to

individual brain patterns (Gu et al., 2022). This approach aims to not only train ANNs to predict brain responses accurately at the individual level, but also capture idiosyncrasies in brain organisation. Exploring this concept further, Seeliger and coll. developed a multilayer convolutional network trained on 40 hours of fMRI data from a single individual watching episodes of the television show Dr. Who (Seeliger et al., 2021), while St-Yves and coll. trained individual convolutional networks with a series of static image presentations on the NSD (St-Yves et al., 2023). The potential of such individualised models is not confined to encapsulating subject-specific brain organisation; there's also evidence of efficient generalisation to downstream tasks when fine tuning models for brain encoding, as demonstrated with electroencephalography data in the vision domain (Palazzo et al., 2020).

Courtois NeuroMod

Despite these advancements towards individual brain models, in particular in the field of vision, we are not aware of attempts to train auditory ANNs with individual brain data and assess its impact on auditory task performance. In this work, we take 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. The Courtois NeuroMod dataset features a wide range of tasks across multiple domains, and includes several movie watching datasets featuring a complex soundtrack. This dataset thus opens the possibility to build individual models of sound processing.

Specific objectives, hypotheses and outline of experiments

This study's main objective was to individually align auditory ANNs with brain data and investigate the impact of this alignment on performance across a range of auditory tasks. We sought to:

1. Align a pretrained, task-optimised ANN with individual brain activity.
2. Examine whether individually aligned ANNs exhibit personal characteristics.
3. Assess whether brain alignment improves performance in downstream tasks.

Taken together, these aims establish the feasibility of building personalised models of auditory processing using deep learning tools.

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 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 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].

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.

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

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. 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 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. 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 level 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.

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

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, (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, 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. As a result, SoundNet representations can encode audio features of varying durations, and with increasing abstraction in deeper layers.

fMRI encoding layer and fixed training parameters

Previous work has shown the feasibility of using transfer learning or fine-tuning of deep learning backbones to train encoding models of fMRI activity (Farrugia et al., 2019; Schwartz et al., 2019). One of the originality of these previous studies was to use a trainable model such as a multi-layer perceptron or a CNN to predict brain activity, instead of a linear model, the ridge regression being the most commonly used, e.g. by Kell and coll. (Kell et al., 2018). In this paper, we input raw audio waveforms resampled at 22050 Hz to extract features using SoundNet, and we train an additional convolutional layer to predict the fMRI signal. Precisely, we explored audio waveforms of varying lengths (from 20 s up to 130 s) and used the output of the seventh convolutional layer of SoundNet (conv7) as input of the encoding convolutional layer. The reason behind the choice of layer Conv7 output is that it has the same temporal resolution as the fMRI signal (0.67 Hz). We implemented the SoundNet architecture followed by the fMRI encoding layer as a fully end-to-end trainable network, by adapting an open-source pytorch implementation of SoundNet[4].

To train this architecture, we used AdamW (Loshchilov & Hutter, 2019) as an optimizer for the L2 regularisation and the weight decay, and we applied a learning rate scheduler that reduces the initial learning rate if no progress is achieved by the optimizer. MSE loss is used as our loss function to minimise the difference between the prediction and the actual fMRI signal. We added a padding to the kernels of the encoding layer (size : Kernel size - 1).

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 only training, and left out 25 percent for validation. Season 4 was left out to be used only for testing.

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).

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

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	$\mathbf{1.10^{-3}}$	**15**	**0**
	70	1.10^{-4}	5			
	80	$\mathbf{1.10^{-5}}$	**6**			
06	60	1.10^{-3}	4	$\mathbf{1.10^{-3}}$	**15**	**0**
	70	1.10^{-4}	5			
	80	$\mathbf{1.10^{-5}}$	**6**			
01	60	1.10^{-4}	5	$\mathbf{1.10^{-3}}$	**15**	**0**
	70	$\mathbf{1.10^{-5}}$	6			
	80	1.10^{-6}	**7**			
02	60	1.10^{-4}	5	$\mathbf{1.10^{-3}}$	**15**	**0**
	70	$\mathbf{1.10^{-5}}$	6			
	80	1.10^{-6}	**7**			
05	**60**	1.10^{-4}	5	$\mathbf{1.10^{-3}}$	**15**	**0**
	70	1.10^{-5}	6			
	80	$\mathbf{1.10^{-6}}$	**7**			

Table 1. Hyperparameters values explored for subjects' baseline model. The values selected for fine-tuning the models are marked in bold. 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 the previous subjects appeared to be the maximal value of the explored window. 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.

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 as the Baseline model. 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, see Table 1 :

- 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. With a range of different values explored for each of the different criteria (see Table 1), we explored 1296 possible configurations, for both whole-brain and middle STG models. As results were equivalent between both models, we decided to only keep results from Whole-brain models. To select the final hyperparameters, we sorted every configuration by the r^2 score obtained on the validation set.
We computed a correlation matrix between ROI predictions from all configurations, and we observed clusters of configurations. To better define these clusters, we used an rch Articles, but we suggest that authors try not to exceed 5,000 words in the main text, excluding the Materials and methods, References, and Figure legends. There are no limits on the number of display items. The main text of the article should usually be structured and ordered as follows: Introduction; Results; Discussion; Materials and Methods (or Methods); Acknowledgements; References; Figures with the corresponding legend below each one; and Tables. A Methods or Model section can appear aagglomerative hierarchical clustering, with a linkage function computing the Euclidean distance between centroid of clusters (UPGMC algorithm, as implemented by the Scipy library[5]). We then determined which value of each parameter was prevalent in the best cluster, by computing the statistical mode for the 20 configurations present in the cluster with the higher r^2 score.

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 1). As a result, we explored only 27 configurations in the remaining subjects. For each subject, 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.

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

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.

Models Comparison and statistical analysis

In order to compare performances between 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 corrected p-value inferior to 0.05 were considered as significant. We used the same procedure to determine if the difference of r^2 scores between baseline and fine-tuned were significant.

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, 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, 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.

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

Results

Baseline brain encoding using pretrained SoundNet

Optimal hyperparameters substantially improve brain encoding with SoundNet

We optimised the hyperparameters for training our brain encoding model, which predicts individual brain activity using SoundNet internal representations. Through a systematic grid search, we explored the impact of several key hyperparameters, including learning rate, temporal window size, convolutional kernel size, weight decay, and early stopping parameters.

We tested 1296 configurations of hyperparameters on the fMRI data from sub-03. When ordering each configuration by the maximal r^2 score obtained on the validation set for the Whole Brain baseline model, we observed an important gap between the first and last models: models with best r^2 scores scored as high as 0.39, when models ranking 400 and below achieved a r^2 score of 0.07 at best. A similar result has been found for the middle STG baseline model, with a maximal r^2 score of 0.45, with a decrease of the score up to 0.05 for the worst performing models. This gap has also been observed when clustering the 100 best models. The agglomerative clustering of the r^2 brain encoding maps using the centroid method with Euclidean distance found 3 clusters of configurations for the Whole Brain baseline model: the mean of the maximal r^2 scores of the best cluster, with 20 configurations, is 0.36, while for the second cluster, the score drops to 0.29 (70 configurations), and to 0.2 for the last cluster (10 configurations).

We used the best performing cluster to determine the prevalent values for each hyperparameter tested. We defined these values as optimal for sub-03, and evaluated a smaller set of hyperparameters values for other subjects, resulting in testing 27 configurations for each subject (refer to methodology for more details). We selected hyperparameters values for every other subject than sub-03 by looking for the prevalent values in the 10 best performing configurations amongst the 27.

Through these experiments, we identified an optimal set of hyperparameters that were then used for fine-tuning the models. Overall, our hyperparameter optimization process played a crucial role in enhancing the quality of brain encoding and selecting the most effective hyperparameter configuration for the subsequent analyses.

SoundNet successfully encodes brain activity in the auditory cortex

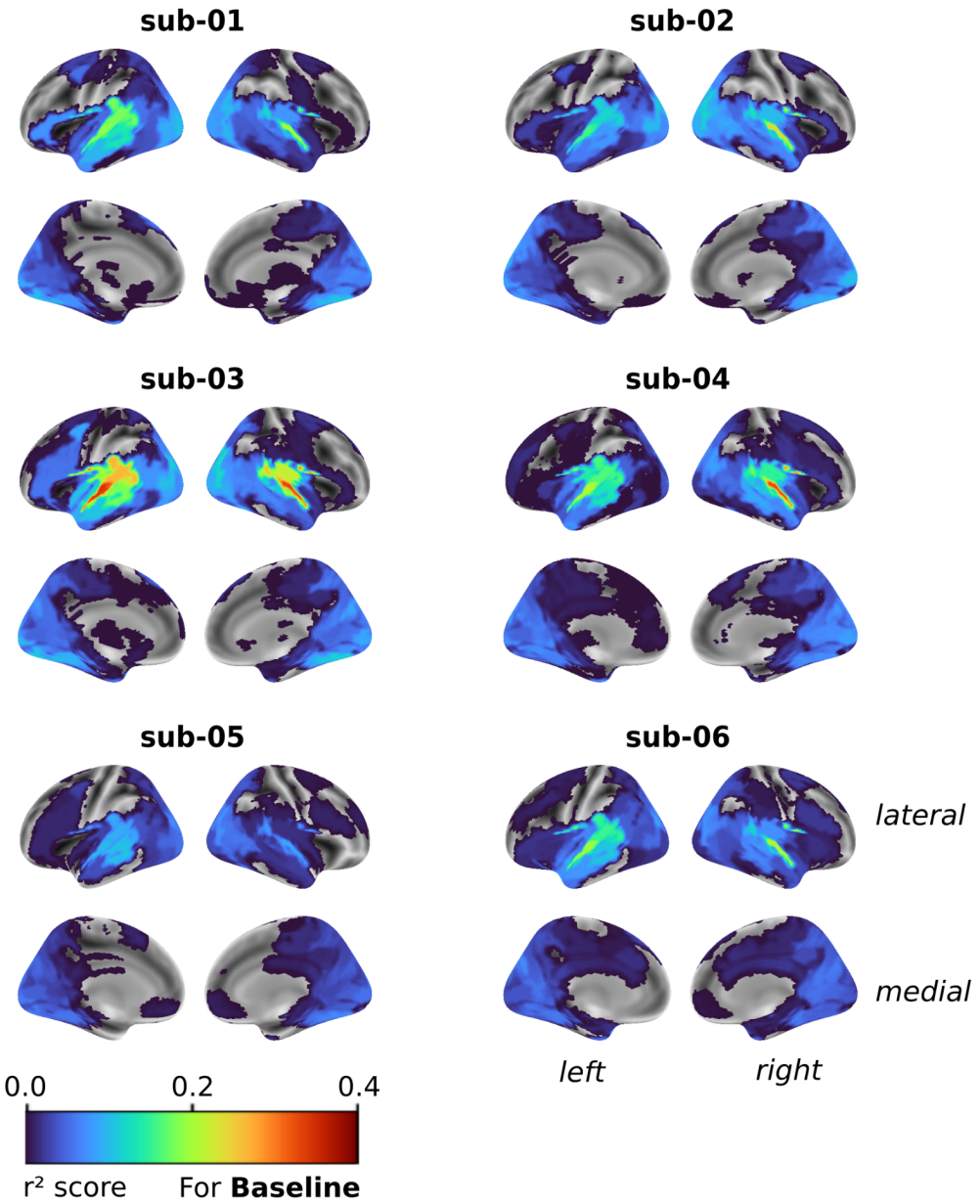

Figure 1. Full brain encoding using SoundNet. Surface maps of each subject, showing the r^2 value for all ROIs from the MIST ROI parcellation. Only parcels with r^2 values significantly higher than those of a null model initialised with random weights are shown (Wilcoxon test, FDR corrected $p < 0.05$). 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. The left and right STGm had some of the highest r^2 scores ($q < 0.05$ for all subjects). The posterior STG (STGp) also consistently ranked second in terms of prediction accuracy (see Supplementary Figure 1). Subject 05 was an exception, with the highest r^2 score in the left Middle Temporal Gyrus (MTG) superior, followed by the STG.

Though the STG was best predicted, 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 successfully encodes activity brain activity in the visual cortex and other regions

Apart from the STG middle and posterior being the best ROI predicted, 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 successfully encodes high resolution brain activity in the Superior Temporal Gyrus

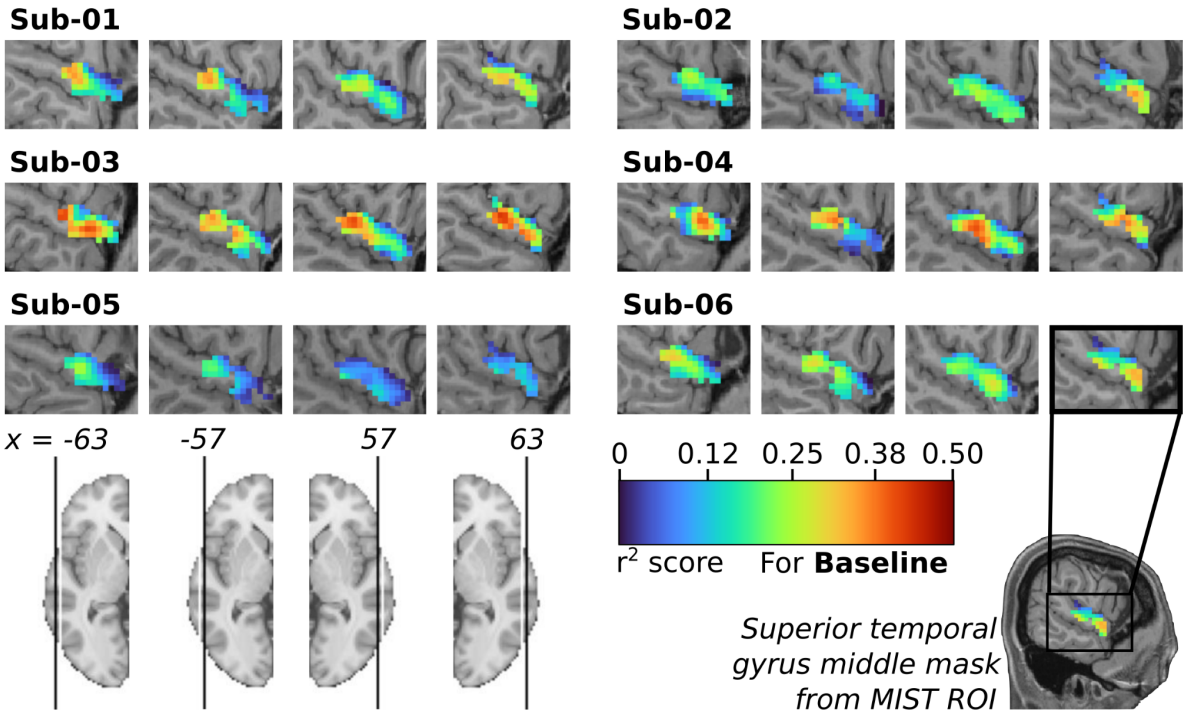

Figure <Fig_STG_baseline>. STG encoding using SoundNet. Mapping of the r² scores from 556 voxels inside the cerebral region defined as the Middle STG by the parcellation MIST ROI, computed by the individual Baseline model. To have a better representation of the STG, 4 slices have been selected in each subject, 2 from the left hemisphere (-63 and -57) and 2 from the right hemisphere (63 and 57). Only voxels with r² values significantly higher than those of a null model initialised with random weights are shown (Wilcoxon test, FDR corrected $p < 0.05$). 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 r² 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 r² of 0.45), while subject 05 performed lower

(average max r² of 0.27). 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

Conv4 fine-tuning leads to substantial improvements in brain encoding

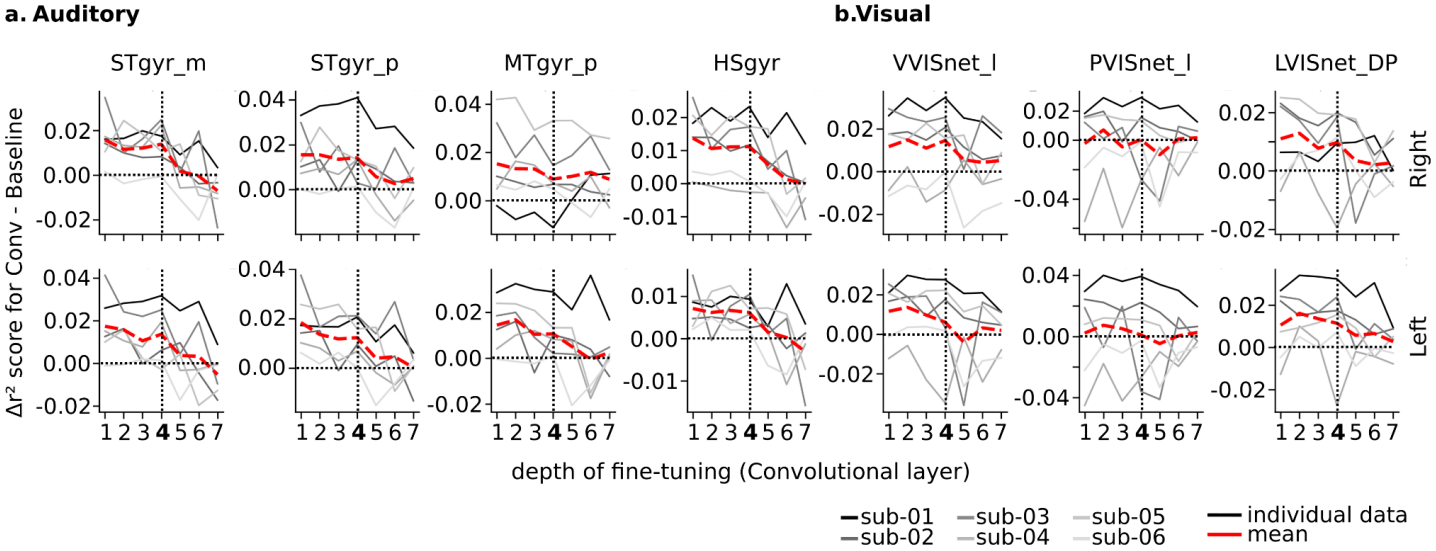

Figure 3. Impact of fine-tuning internal layers in SoundNet on individual brain encoding. Difference between the r² score (median across test runs) of the fine-tuned models (Conv1 to Conv7) minus baseline for each subject (sub-01 to sub-06) and the mean. Graphs are presented in the best predicted auditory (panel a) and visual (panel b) regions for the baseline model. STgyr_m: Superior Temporal gyrus middle; STgyr_p: Superior Temporal gyrus posterior; MTgyr_p: Middle Temporal gyrus posterior; HSgyr: Heschl's gyrus; VVISnet_l: Ventral Visual network lateral; PVISnet_l: Posterior Visual network lateral; LVISnet_DP: Lateral Visual network dorsoposterior.

Our main goal was to see if fine-tuning an auditory model improved brain encoding performance. We adjusted SoundNet at different depths, using our training data (seasons 1 to 3), then tested the model using season 4. We calculated the median r² score for each of the 210 parcels of the MIST ROI atlas and averaged the performance across all subjects (Figure 3).

The best-predicted ROIs by the baseline model showed a consistent pattern (Figure 3). Adjusting models at Conv7, Conv6, or Conv5 did not improve brain encoding and even worsened it for the STG middle. However, models adjusted at Conv1 to Conv4 significantly improved over the baseline, with Conv4 showing the most substantial improvements. Adjusting at Conv1 to Conv3 only gave slight or no improvements over Conv4 for most ROIs, except the left lateral posterior visual network. We decided to focus on Conv4 models for further experiments based on these results.

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

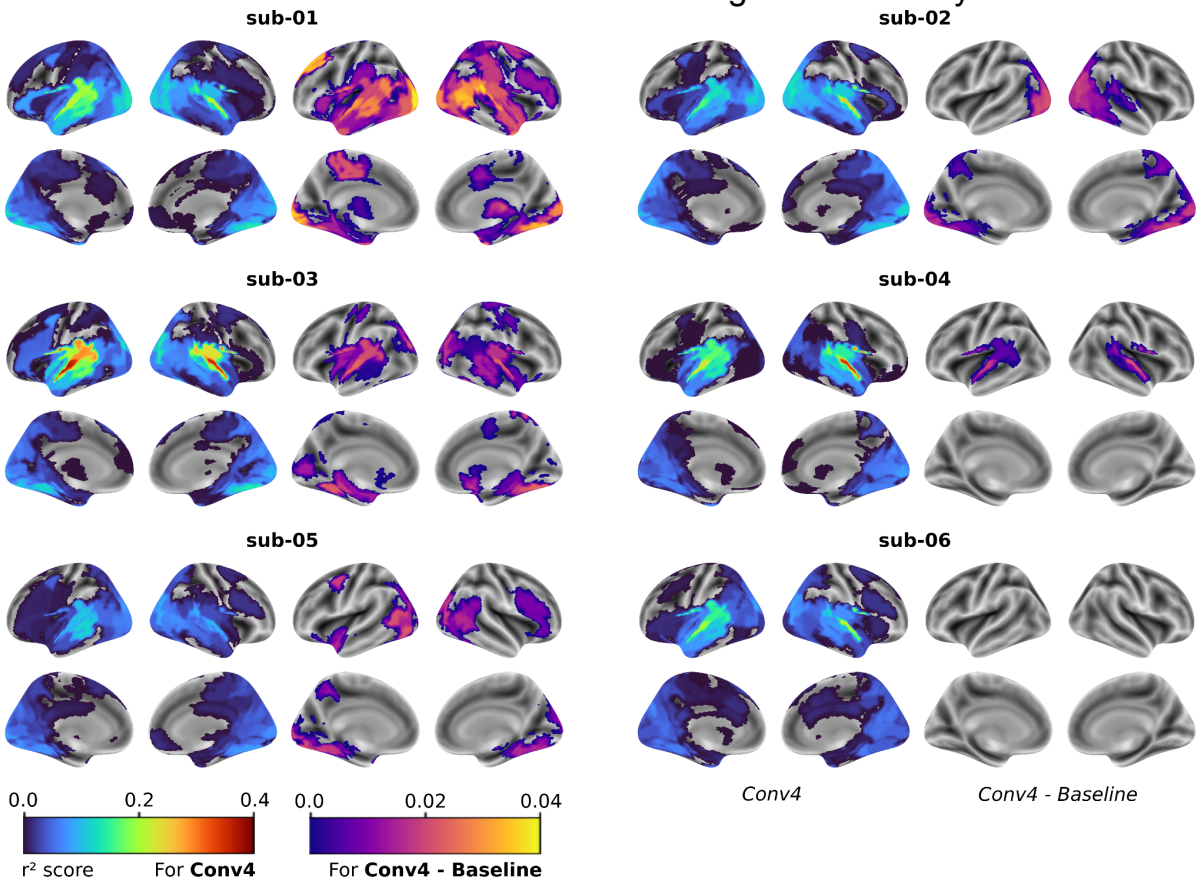

Figure 4. Individual impact of Conv4 fine-tuning on the full brain encoding. For each subject, on the left side : Surface maps of the r² scores computed with each individual Conv4 model, for the 210 ROIs of the MIST ROI parcellation. Coloured ROIs have a r² score significantly greater than the null model ($p<0.05$). On the right side : surface maps of the difference in r² scores computed with each individual Conv4 and Baseline models, for the 210 ROIs of the MIST ROI parcellation. Coloured ROIs have a Conv4 r² score significantly greater than the baseline model ($p<0.05$)

We next examined which brain ROIs had the most improvement using Conv4 fine-tuned models. We tested each ROI's r² score for significant improvements over the baseline and created a brain map (Figure 4). The top-predicted parcels for Conv4 models were the same as those at baseline. For most subjects, STGm and STGp remained the highest-scoring ROIs. However, for subject 02, some visual cortices were best encoded, with STG regions as a close second.

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 r² scores compared to the null model. In contrast, models fine-tuned with subjects 03 and 04's data showed fewer significant ROIs. Subject 06's model showed no improvements from fine-tuning.

The fine-tuning's main improvements were not always in the auditory cortex. We observed r² score improvements in the visual cortex, especially in the lateral and posterior visual network, for subjects 01, 02, 05, and to a lesser extent, subject 03. No improvements were seen in subject 04, where the only improvements were in the primary and secondary auditory cortex, and the model's predicting capacity in the frontal and parietal cortex was degraded by fine-tuning. In the case of subject 06, the fine-tuning had no impact at all.

Overall, the 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

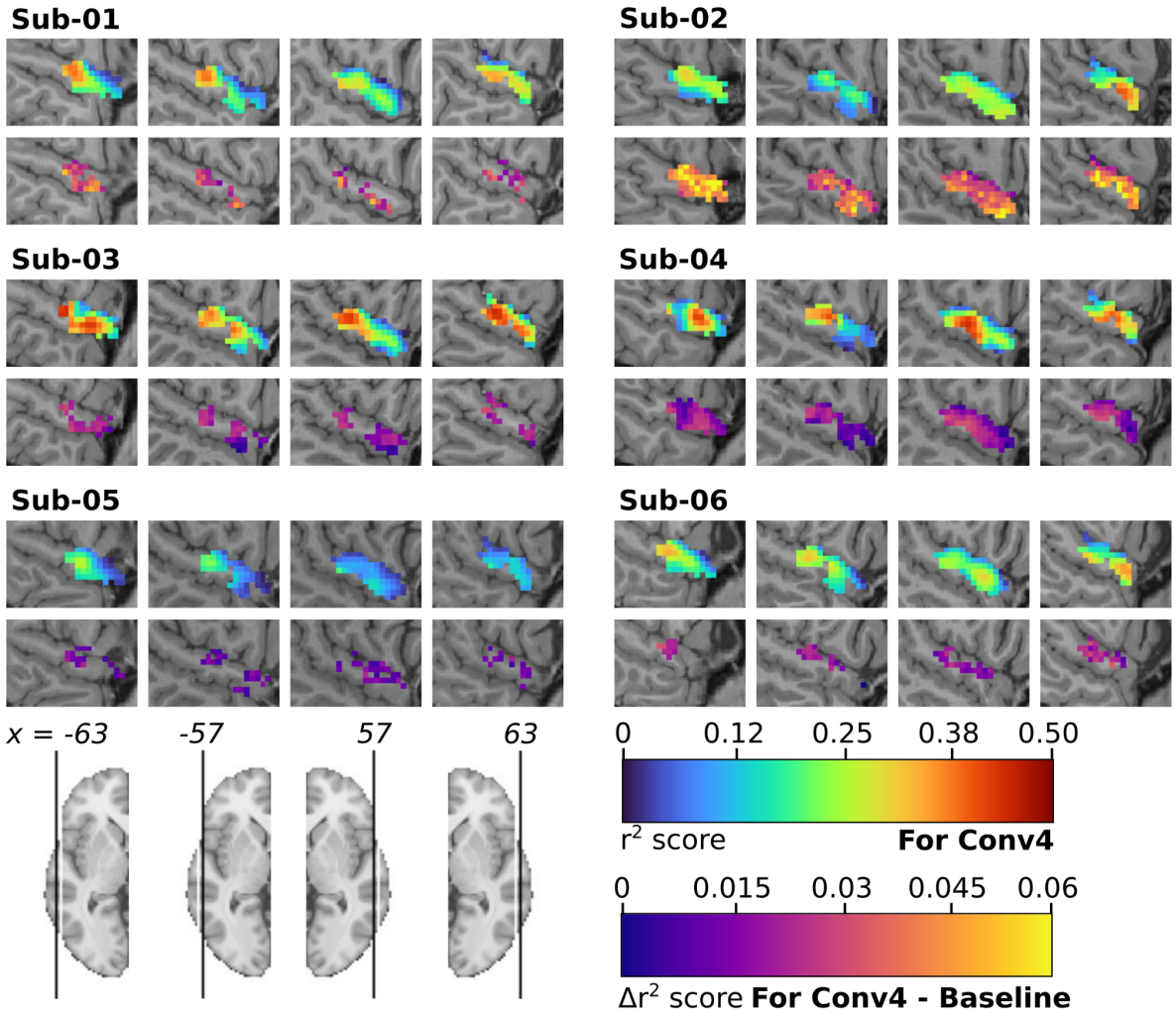

Figure 5. STG encoding using SoundNet. For each subject, on the top : mapping of the r² scores from 556 voxels inside the cerebral region defined as the Middle STG by the parcellation MIST ROI, computed by the individual Conv4 model. Only voxels with r² values significantly higher than those of a null model initialised with random weights are shown (Wilcoxon test, FDR corrected $p < 0.05$). For each subject, on the bottom : mapping of the difference of r² scores between the Conv4 model and the baseline model. Only voxels from the Conv4 model with r² values significantly higher than those of the baseline model are shown (Wilcoxon test, FDR corrected $p < 0.05$). 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 5). 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. The fine-tuning improved subject 05's model less than others, as indicated by a smaller general r² score increase. These findings align with whole brain model results. Interestingly,

subject 06's Conv4 model showed improvements in the left and right STG middle after fine-tuning, contrasting with whole brain model findings. 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-tuned Conv4 models are subject specific

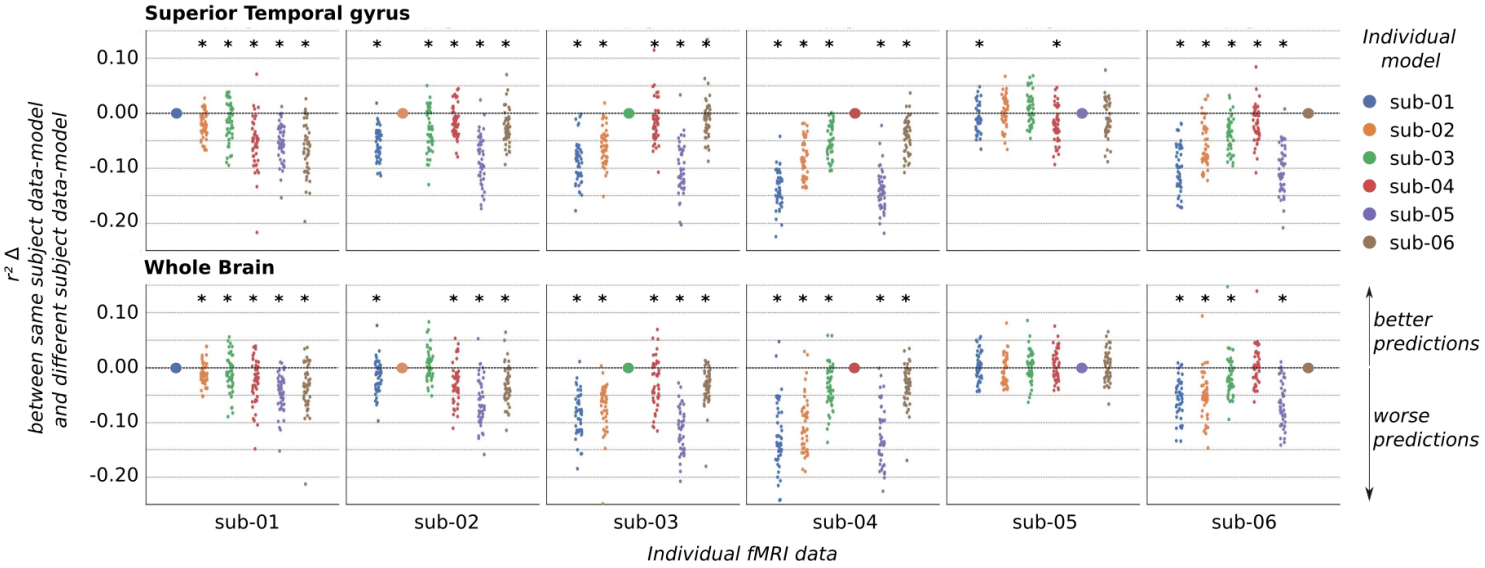

Figure 6. Difference of r² scores computed by an individual Conv4 model, predicting fMRI activity from the same subject that the data used to train it, versus predicting fMRI activity from another subject. The difference is computed for each of the 48 runs of the fourth season of Friends. A Wilcoxon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

We evaluated next 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.

Fine-tuning improves SoundNet ranking in diverse AI auditory benchmarks

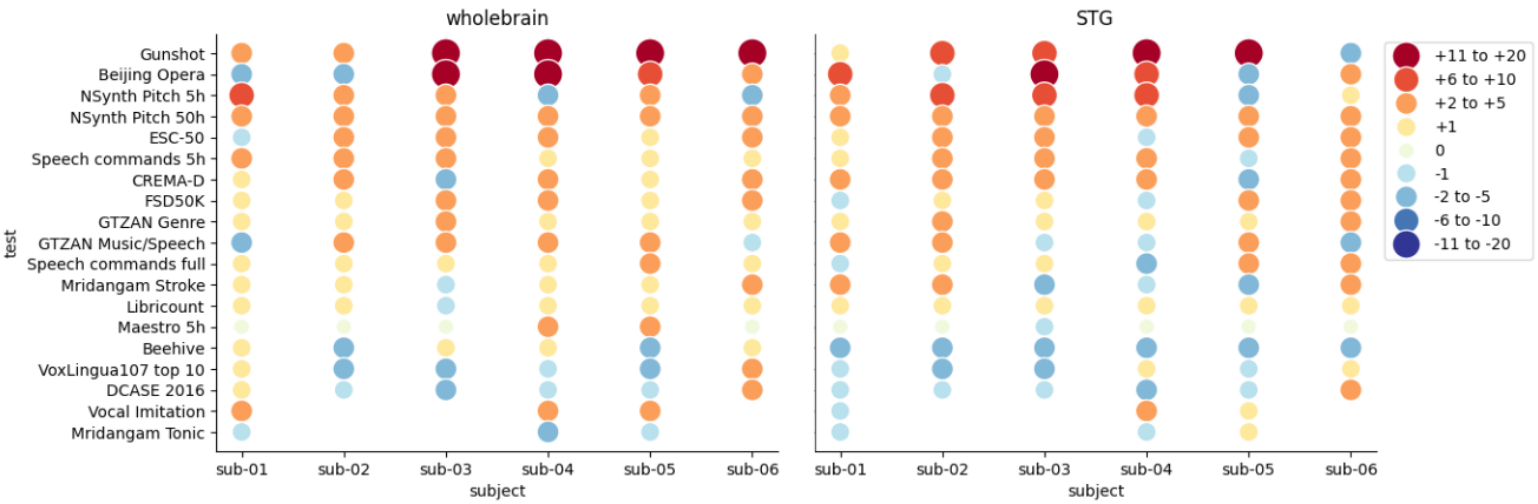

Figure 7. Rank variation between Conv4 and baseline models on all tasks from the HEAR benchmark. Each individual Conv4 model (both Whole brain and middle STG models) has been used to resolve the 19 tasks from the HEAR Benchmark. We followed the directives given by the organisation so that our models can be used with the HEAR API; results from both baseline and Conv4 models are compared to the scores from 29 other models, publicly available on the HEAR website. 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 Baseline at the task, but doesn't outperform other models already superior to Baseline.

We finally 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 each fine-tuned model against SoundNet, and ranked both of them amongst the 29 models which had been tested with this benchmark. All 6 individual fine-tuned models performed significantly better than SoundNet ($p < 0.05$) (left panel of the figure 7), using Whole Brain fine tuning. Results with individual middle STG models are more heterogeneous, with only sub-02, sub-03 and sub-06 Conv4 models ranking significantly higher than SoundNet.

Fine-tuning did not improve the performance of the model in all tasks: some specific tasks seem to benefit much more from the finetuning 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. For most subjects, Whole brain models show a higher mean gain in rank than the STG models, with the exception of sub-02 models. We noted that the Whole Brain fine-tuned Sub-06 model seemed to improve performance in numerous tasks, while no improvement has been found in the brain encoding aspect.

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

Discussion

In this study, we have explored the benefits of aligning a pretrained auditory neural network (SoundNet) to individual brain activity. Our results confirm considerable improvements in encoding brain activity, with gains extending beyond the auditory cortex. Importantly, we found that these brain-aligned, individual-specific networks were more accurate in predicting brain activity for the subjects on which they were trained, compared to the performance of models trained on data from other subjects. Furthermore, 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.

Can task-optimized ANN be aligned with brain activity?

Our findings suggest task-optimized ANNs can successfully be aligned with individual brain activity. Specifically, the fine-tuning process we implemented resulted in substantial enhancements in brain encoding quality. The hyperparameters grid search revealed the importance of learning rate, size of the kernels in the convolution layer as well as the length of the input to have a major impact on the prediction. We also found that fine-tuning the top half of SoundNet (conv4 layer and above) had the most impact on the quality of brain encoding. The extent of these improvements varied across both brain regions and individuals. All participants exhibited significant advances in the superior temporal gyrus (STG), with increases in R2 scores of up to 0.06, when models were directly fine-tuned to encode voxel-level STG activity. Furthermore, four out of six individuals showed considerable enhancements (up to 0.04 in R2 scores) in the STG region when models were fine-tuned on the entire brain. Several regions outside the STG, such as the visual cortex, also experienced improvements comparable to those in the STG. Given the known activation of visual cortices by auditory stimuli, as well as the correlation of audio and visual features in our video stimuli (for instance, the presence of faces and lip movements during speech), this was not unexpected. However, 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.

Are brain-aligned ANNs learning subject-specific representations?

Do brain-aligned ANNs learn subject-specific representations? We found that models fine-tuned to individual subjects' fMRI data did not translate well to other subjects. Both whole-brain and STG models were more accurate when predicting data from their training subjects, with STG models showing higher individual specificity. This aligns with studies by Schwartz and collaborators (Schwartz et al., 2019) and Rastegarnia and collaborators (Rastegarnia et al., 2022) which also discovered idiosyncratic, subject-specific brain representations. Similarly, Porter and collaborators (Porter et al., 2023) reported that individual models significantly outperformed a group model in a task-based decoding study. Our results echo these findings, suggesting a reliable and unique network organisation at the individual level in deep phenotyping datasets. However, our research only tested individual specificity without exploring potential alignment of representations across subjects. Future work could investigate this by using hyper-alignment methods (Bazeille et al., 2021) to decipher which components of the model's auditory latent representations are genuinely subject-specific versus those simply mapping to varying cortical territories.

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. Overall, the models aligned to whole-brain data outperformed those aligned to the middle STG. However, task-specific performance varied significantly, with some tasks, such as Gunshot triangulation, showing substantial improvements from whole-brain models. Conversely, tasks requiring pitch differentiation, like Nsynth Pitch, benefited more from middle STG models. This variability extended to individual subjects, as models demonstrated different performance levels across tasks. Interestingly, our models' representations seem to be impacted by fine-tuning even without significant changes in brain encoding scores. To understand these results, further analysis of the models' internal representations is needed.

Only two previous studies have employed a large benchmark after brain data fine-tuning. The Palazzo study (Palazzo et al., 2020) reported marginal performance gains in vision tasks post-alignment with EEG data, while the Schwartz study (Schwartz et al., 2019) reported no significant change in performance after brain alignment with fMRI or MEG data. Our research differs notably in auditory processing modality, stimulus nature (a TV show), and the data volume used for fine-tuning. We recognize the limited sample size used in this study and aim to expand future research. The forthcoming Courtois NeuroMod sample will cover a range of stimulus types and modalities, allowing for further comparative studies on fine-tuned models. Additionally, we're interested in investigating whether brain-aligned models lead to human-like similarity judgement patterns (Bakker et al., 2022).

Limitations

Our study was limited by its exclusive use of the SoundNet pretrained convolutional network. Other models we initially tested didn't yield satisfactory brain encoding results, hence our selection of SoundNet. However, these preliminary analyses lacked systematic hyper-parameter tuning and were conducted on varied datasets. Since our selection in 2019, AI and brain encoding research fields have advanced considerably, with more recent DNNs surpassing SoundNet's performance across all HEAR benchmark audio tasks.

Though CNNs are still popular in brain encoding and decoding studies, particularly in vision due to convolution's similarity to early visual layer processing (Seeliger et al., 2021), Natural Language Processing networks like BERT and transformers have shown promise (Conwell et al., 2021; Schwartz et al., 2019). Transformers have been effective for brain encoding in vision and multimodal (language and vision) domains. However, the time required to develop an end-to-end learning framework and conduct a hyperparameters grid-search has precluded the inclusion of other architectures in this study.

Future work should compare our results with studies using different models for auditory processing brain encoding. With the rise of paradigms for rapid fine-tuning of large pretrained networks, such as LoRA (Hu et al., 2021), this could be expedited. We also plan to test our models' generalisation capabilities on a variety of audio stimuli, as the Courtois NeuroMod project is collecting a broad array of task stimuli from the same six individuals.

Impact

Our study demonstrates that large-scale neuroimaging data gathered using naturalistic stimuli can be utilised to develop individual models of auditory processing capable of complex behaviour. These individual brain-aligned models unlock numerous research opportunities at the intersection of neuroscience and AI, such as examining brain representation idiosyncrasies, assessing representation generalizability to different stimuli types, and investigating how fine-tuning affects different model architectures or training data.

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 improved performance on an extensive benchmark of downstream AI audio tasks. Individual brain models also did not generalise to other individuals, implying idiosyncrasies in either the learned representations or the brain mapping layer. These findings open a wealth of opportunities for future research, ranging from studying inter-individual variations to testing various model architectures and types of training data.

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".

Authors Contributions

Prototyping, development, and finetuning 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.
Manuscrit has been written by MF and PB, with additional reviewing and contributions on the methods section by NF.

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.

Aytar, Y., Vondrick, C., & Torralba, A. (2016). *SoundNet: Learning Sound Representations from Unlabeled Video*. arXiv. https://doi.org/10.48550/arXiv.1610.09001

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

Bazeille, T., DuPre, E., Richard, H., Poline, J.-B., & Thirion, B. (2021). An empirical evaluation of functional alignment using inter-subject decoding. *NeuroImage*, *245*, 118683.

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.

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.

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?* https://openreview.net/forum?id=i_xiyGq6FNT

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.

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). *Estimating encoding models of cortical auditory processing using naturalistic stimuli and transfer learning*. https://openreview.net/forum?id=SyxENQtL8H

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.

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.

Gu, Z., Jamison, K., Sabuncu, M., & Kuceyeski, A. (2022). Personalized visual encoding model construction with small data. *Communications Biology*, *5*(1), 1–12.

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.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). *LoRA: Low-Rank Adaptation of Large Language Models*. arXiv.

https://doi.org/10.48550/arXiv.2106.09685

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*. https://papers.nips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

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

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.

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

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

Poldrack, R. A., Laumann, T. O., Koyejo, O., Gregory, B., Hover, A., Chen, M.-Y., Gorgolewski, K. J., Luci, J., Joo, S. J., Boyd, R. L., Hunicke-Smith, S., Simpson, Z. B., Caven, T., Sochat, V., Shine, J. M., Gordon, E., Snyder, A. Z., Adeyemo, B., Petersen, S. E., … Mumford, J. A. (2015). Long-term neural and physiological phenotyping of a single human. *Nature Communications*, *6*(1), 8885.

Porter, A., Nielsen, A., Dorn, M., Dworetsky, A., Edmonds, D., & Gratton, C. (2023). Masked features of task states found in individual brain networks. *Cerebral Cortex* , *33*(6), 2879–2900.

Rastegarnia, S., St-Laurent, M., DuPre, E., Pinsard, B., & Bellec, P. (2022). *Brain decoding of the Human Connectome Project Tasks in a Dense Individual fMRI Dataset*. PsyArXiv. https://doi.org/10.31234/osf.io/9t5nh

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*. https://proceedings.neurips.cc/paper/2019/hash/2b8501af7b64d1aaae7dd832805f0709-Abstract.html

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., Malloy, C., Tzanetakis, G., Velarde, G., McNally, K., Henry, M., Pinto, N., Noufi, C., Clough, C., Herremans, D., Fonseca, E., Engel, J., Salamon, J., Esling, P., … Bisk, Y. (2022). *HEAR: Holistic Evaluation of Audio Representations*. arXiv. https://doi.org/10.48550/arXiv.2203.03022

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.

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

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.