Summer 8-14-2023

Self-Supervised Pretraining and Transfer Learning on fMRI Data with Transformers

Sean Paulsen
Dartmouth College, paulsen.sean@gmail.com

SELF-SUPERVISED PRETRAINING AND TRANSFER LEARNING ON FMRI DATA WITH TRANSFORMERS

A Thesis
Submitted to the Faculty
in partial fulfillment of the requirements for the
degree of

Doctor of Philosophy

in

Computer Science

by Sean Paulsen

Guarini School of Graduate and Advanced Studies
Dartmouth College
Hanover, New Hampshire

August 2023

Examining Committee:

Michael Casey, Chair

Soroush Vosoughi

SouYoung Jin

Petr Janata

Tor Wager

F. Jon Kull, Ph.D.
Dean of the Guarini School of Graduate and Advanced Studies

Abstract

Transfer learning is a machine learning technique founded on the idea that knowledge acquired by a model during "pretraining" on a source task can be transferred to the learning of a target task. Successful transfer learning can result in improved performance, faster convergence, and reduced demand for data. This technique is particularly desirable for the task of brain decoding in the domain of functional magnetic resonance imaging (fMRI), wherein even the most modern machine learning methods can struggle to decode labelled features of brain images. This challenge is due to the highly complex underlying signal, physical and neurological differences between brains, low data collection throughput, and other factors. Transfer learning is exciting in its potential to mitigate these challenges, but with this application still in its infancy, we must begin on the ground floor.

The goals of this thesis were to design, implement, and evaluate a framework for pretraining and transfer learning on arbitrary fMRI datasets, then demonstrate its performance with respect to the literature, and achieve substantive progress toward generalized pretrained models of the brain. The primary contribution is our novel framework which achieves these goals, called BEAT, which stands for Bi-directional Encoders for Auditory Tasks. The design and implementation of BEAT include adapting state-of-the-art deep learning architectures to sequences of fMRI data, as well as a novel self-supervised pretraining task called Next Thought Prediction and several novel supervised brain decoding tasks. To evaluate BEAT, we pretrained

on Next Thought Prediction and performed transfer learning to the brain decoding tasks, which are specific to one of three fMRI datasets. To demonstrate significant benefits of transfer learning, BEAT decoded instrumental timbre from one of the fMRI datasets which standard methods failed to decode in addition to improved downstream performance. Toward generalized pretrained models of the brain, BEAT learned Next Thought Prediction on one fMRI dataset, and then successfully transferred that learning to a supervised brain decoding task on an entirely distinct dataset, with different participants and stimuli. To our knowledge this is the first instance of transfer learning across participants and stimuli–a necessity for whole-brain pretrained models.

Acknowledgements

First and foremost I simply would not have ever pursued a PhD without the excellent educators from whom I have had the unimaginable good fortune to learn. Adam LaMee, my high school physics teacher, showed me and everyone else how exciting science can be and changed the course of my life more than any other individual – I wanted to be a lawyer before I stepped into his class. Katherine Spradlin taught my AP English class, but much more importantly was the only teacher willing to consistently check the ego of an annoying know-it-all who thought he was above it all. Stephen Fannin was my high school chemistry teacher, who taught me as much about teaching as he did about chemistry. I have quoted his advice directly to lock down more than one job interview over the years.

I never would have been able to cross over to computer science from mathematics without the masterful instruction of Chrysafis Vogiatzis in C++ at the University of Florida. He also hired me as an undergraduate teaching assistant, allowing me to discover my love of and talent for teaching. Michael Jury was my professor at UF for multiple math courses and supervised my undergraduate thesis. My academic excellence at Dartmouth is almost entirely due to the way he taught me to conceptualize and internalize complex ideas. I cannot possibly overstate his impact on my success.

I want to thank Michael Casey for his excellent guidance as my PI and thesis advisor at Dartmouth, as well as his incredible patience and understanding when I felt that I did not deserve it. I want to thank Vasanta Kommineni for allowing me

to thrive as a Teaching Assistant at Dartmouth, and for submitting me for the CS Department's Best TA Award. Soroush Vosoughi, Andrew Campbell, Souyoung Jin, Tor Wager, and Petr Janata have my insufficient gratitude for generously serving on my thesis committee.

Jason Burkhart has been my closest friend for more than half my life and I would be lost without him. In no particular order, Andy Downing, Arie Brown, Brendan Faeth, and my parents have all supported me as much as possible in their own ways and I will be grateful forever.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

"This it is necessary to grasp, but not easy."

–Aristotle, *De Anima*

This chapter provides the introduction to this thesis. In Section 1.1 we give a general overview of the motivations and central concepts, including related works. In Section 1.2 we briefly present our main contributions.

Section 1.1

General Introduction

Functional MRI (fMRI) measures blood-oxygen-level-dependent (BOLD) responses that reflect changes in metabolic demand consequent to neural activity[7, 38, 91]. By measuring BOLD responses at specific combinations of spatio-temporal resolutions and coverages, fMRI data provide the means to study complex cognitive processes in the human brain[48, 117, 78]. In particular, task-based fMRI protocols include targeted stimuli or other task variables, such as question answering, during the scan. Researchers can then conclude associations between task features and the evoked responses across the brain[54, 114, 71]. Regions of activity that are correlated with the presence of

a particular task feature are thus taken to be involved in the brain's representation of that feature[101], and they are considered to be functionally connected[93]. Even rest-state fMRI data, that is, data collected in the absence of external stimuli or task, contain characteristic multi-variate signals of the brain[79, 64, 72, 126, 111, 42]. Such rest-state signals have been shown to be predictive of the diagnosis and characterization of multiple neurological diseases and psychiatric conditions[130, 120, 124].

fMRI researchers have adopted several data analysis techniques to analyze the complex relationship between BOLD signal and the underlying task, disease, or biological information. More specifically, the task of predicting such information given the BOLD data as input is known as **task-state decoding**, or **brain decoding**. Toward the goal of more powerful brain decoding models, many advances in modern *deep* machine learning have been applied to fMRI research. These include convolution-based models[132, 79], recurrent neural networks (RNN)[14], and graph neural networks[58]. Most recently, Transformer[113] based models have achieved state of the art results on several brain decoding tasks[62, 7, 70], having already grown to dominate the fields of time series forecasting[56], natural language processing[17], and computer vision[19, 55]. Indeed, BOLD signal has high spatio-temporal complexity, so this shift in the literature toward brain decoding on sequences of whole regions of fMRI data, rather than single images or single voxels, has been and continues to be motivated by a desire to capture the temporal as well as spatial components of the signal.

However, training deep models is data intensive, while fMRI scans are expensive with relatively little data obtained per scan. Moreover, the experiment-specific labels and scanning parameters of the images mostly eliminate the possibility of combining datasets. In Paulsen, May, and Casey (2021)[79], our work prior to beginning this thesis, we employed a deep learning model to learn the latent patterns in short sequences of *unlabelled* fMRI data, then leveraged the knowledge of those patterns to

improve (and enable) success of downstream linear classifiers. This strategy can be thought of as an early form of **sequential transfer learning**, in which a model is first **pre-trained** on a task in the domain of interest to acquire general knowledge about the target dataset. The pretrained model then has a head start, so to speak, on the target task of interest, by leveraging its general understanding of the data[21]–in other words, by transferring its learning. As Kalyan et al. (2021)[45] note about pretrained models, "These models provide good background knowledge to downstream tasks which avoids training of downstream models from scratch." This strategy is nearly ubiquitous in the domain of Natural Language Processing (NLP)[45] and has begun to appear in fMRI studies aside from our own[70, 62, 7]. We discuss our previous work[79] in more detail in the next chapter, but its success was our primary motivator for developing our own modern framework for transfer learning with fMRI data.

In 2019, Devlin et al.[17] presented BERT, which stands for Bidirectional Encoder Representations from Transformers. The Transformer[113] is a deep learning architecture which has come to completely dominate sequence-based machine learning and is particularly well suited for transfer learning–we discuss the transformer architecture in detail in the beginning of Chapter 4. BERT was pretrained on a massive corpus of unlabelled text and the authors performed transfer learning to obtain state-of-the-art results on eleven natural language processing tasks. We take several notes of inspiration from BERT, which are discussed in Chapter 4, but BERT also inspires the idea of a generalized pretrained model of the *brain*, that could transfer its learning to a wide variety of brain decoding tasks on various datasets. This is a gargantuan and long-term goal, but in this thesis we present statistically significant evidence of, to the best of our knowledge, the first step towards it.

We call our framework BEAT, which stands for Bidirectional Encoders for Auditory Tasks. "Auditory" because the datasets used in this work are audio-evoked fMRI

data, but the architecture and task design can be applied to an arbitrary fMRI dataset. When we say "framework" throughout this thesis, we refer to the full suite of architecture design and implementation, pretraining and target task design, dataset construction for training these tasks, and evaluation of transfer learning effects. The components of BEAT are thus the contributions of this thesis.

Section 1.2

Main Contributions

The main contributions of this thesis are the following:

1) A paired-sequence Transformer-based architecture for brain decoding tasks. This architecture is detailed in Section 4.2.

2) An audio-evoked fMRI dataset with condition labels for task-state decoding studies and experiments. The motivation for and collection of this dataset are detailed in Section 3.2.

3) Self-supervised pretraining tasks on fMRI data which enable transfer learning. These tasks are detailed in Section 4.3, and the corresponding experiments and results are detailed throughout Chapter 5.

4) Supervised brain-decoding tasks which demonstrate statistically significant benefits from transfer learning. These tasks are detailed in Section 4.4, and the corresponding experiments and results are detailed throughout Chapter 5. We achieve a new level of granularity in the decoding of audio-evoked fMRI data. This achievement is presented in detail in Section 5.2.3. We also demonstrate an enculturation effect in Nucleus Accumbens after a week of exposure to an unfamiliar musical grammar, as well as, to the best of our knowledge, the first

significant effects of transfer learning on a brain decoding task when the pre-training and finetuning datasets are wholly distinct. These two achievements are presented in detail in Section 5.3.

Chapter 2

Brain Decoding – History and Motivation

In this chapter we consider the motivations behind decoding information from the brain and explore the relevant timeline of research. This timeline will reveal the challenges and shortcomings of such research, ultimately leading to the motivations for this thesis. There are several brain data collection modalities that can be used for brain decoding, but in this work we focus on fMRI. The science behind fMRI is detailed in Chapter 3 but we present a general idea below. Thus the aim of this chapter is to set the stage for the goals of this thesis and provide context for our contributions.

Section 2.1

General Motivations

The relationship between physical neurological states and cognition has been a curiosity throughout human history. The term **neural correlates** is used to refer to brain activity that corresponds with and is necessary to produce a particular experience. In *De Anima*, Aristotle considers whether all psychological states are also material states

of the body, commenting "This it is necessary to grasp, but not easy." We certainly reject the notion that psychological states could be *immaterial*, but what remains is perhaps the next steps in Aristotle's consideration: Which neural correlates are identifiable via physical measurements? How are they represented in the brain? What techniques should be used for measurement and identification?

But why do we care about neural correlates? The universal and fundamental nature of these questions allows sheer curiosity as one motivator. Consciousness, for example, is a question as old as itself. In 2019, Nani et al.[69] published a through survey of the neural correlates of consciousness and attention. From their conclusions, "Consciousness has the function of creating a continuous and coherent picture of reality, while attention has the function of attributing relevance to the objects of thought. Consciousness develops along two dimensions, that of wakefulness and that of contents. It can also be conceptually distinguished between phenomenal consciousness (how the world appears to us) and access consciousness (when contents are more or less vivid, intense, and available for focal awareness)."

There are more tangible motivations as well. Every unraveled neural correlate could provide a new vector to diagnosing neurological conditions in a non-invasive way. For example, resting state (no stimuli) fMRI identified significant disruptions in default mode network's co-activation in patients with Alzheimer's disease, and lowered activation has been identified in the supplementary motor area during movement in patients with Parkinson's disease versus controls[70]. Neural correlates can also lead to new treatments. Neurologic Music Therapy, for example, uses the perception of auditory structures and patterns in music as cues to retrain brain function[107]. Although the number of studies and extent of available evidence is greatest in stroke and dementia, there is also evidence for the effects of music-based interventions on supporting cognition, motor function, or emotional wellbeing in people with Parkinson's

disease, epilepsy, or multiple sclerosis[100].

Brain-computer interface (BCI) technology necessarily requires a robust understanding of neural correlates. In 2004, Goebel[29] presented work where two participants were able to play the video game "Pong" against each other in real time while lying in fMRI scanners. At time of writing, Neuralink has recently obtained FDA approval to begin human trials of a fully implantable BCI with the stated mission to "Create a generalized brain interface to restore autonomy to those with unmet medical needs today and unlock human potential tomorrow."

It is clear that the benefits of brain decoding research are extensive, reaching from answering fundamental questions of human existence to improving the everyday lives of people around the world. With these motivations in mind, the next section discusses the techniques involved in and general history of the research into decoding fMRI data.

Section 2.2

Classical Statistical Inference

Functional magnetic resonance imagine (fMRI) exploits blood-oxygen level-dependent (**BOLD**) contrasts to map neural activity associated with a variety of brain functions including sensory processing, motor control, and cognitive and emotional functions. BOLD signal changes are modulated by changes in blood flow due to the metabolic demand of neural activity. A typical fMRI database contains a timeseries of 3D BOLD signal measured at many **voxels** in the brain. A voxel is a 3D cuboid of brain volume, whose dimensions are on the order of millimeters. Thus the task of decoding cognitive function from fMRI data requires analyzing the BOLD signal at each voxel across the time series (Figure 2.1. We use "voxel data" as a shorthand for such a timeseries throughout this thesis. The subsections below provide an overview of the history of

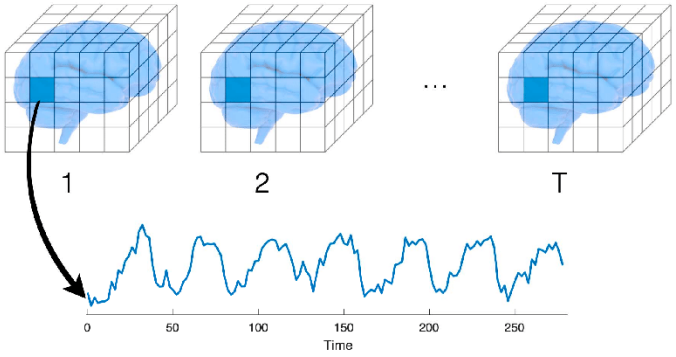

Figure 2.1: The sequences of voxel data used in our experiments are timeseries of neural activity measured by fMRI. Graphic published in [99]

techniques for decoding information from the brain via voxel data analysis.

2.2.1. General Linear Model

The general linear model (**GLM**) approach is used to reveal "activated" brain areas by searching for linear correlations between the fMRI time series and a reference model of the brain defined by the researcher, or with some known pattern of stimulation or experimental manipulation. Iterative statistical analysis on all voxels can then identify regions of voxels whose measured BOLD responses display significant statistical effects. This framework is commonly referred to as **mass-univariate model-based analysis,** and was, at least in 2012, considered the gold standard in fMRI research[34].

The GLM is expressed as a matrix by

$$Y = X\beta + \epsilon, \tag{2.1}$$

where $Y = [y_1, \ldots, y_M]^T$ is a column vector of the recorded BOLD signal at a *single voxel* at each of M-many timesteps. $\beta = [\beta_1 \ldots \beta_P]^T$ is a column vector of unknown model parameters to be estimated. X is the MxP design matrix where the columns are hand-crafted explanatory variables or **regressors** quantifying the experimental knowledge about the expected signal, with the i^{th} row corresponding to the i^{th} timestep.

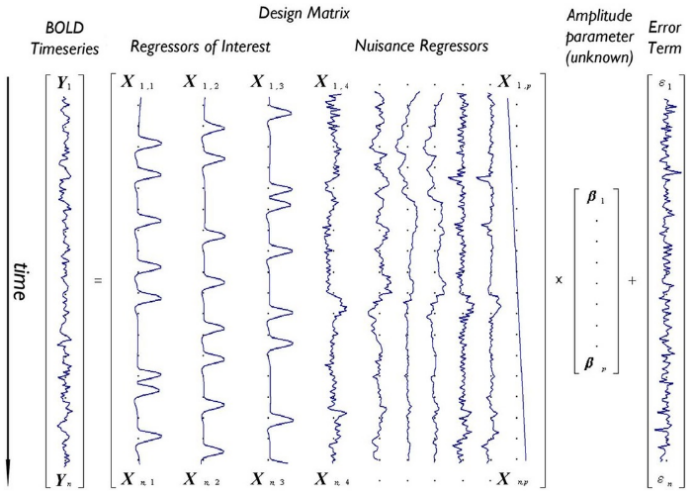

Figure 2.2: General Linear Model (GLM) for a single voxel with timeseries Y, with design matrix X. X has three regressors of interest, corresponding to the conditions during the scan, and seven nuisance regressors to account for confounds such as head motion or signal drifts. Each regressor is weighted by parameter β_i. ϵ_i is the calculated error term at each timestep i. Image from Monti (2011)[66].

β_i can thus be interpreted as the weight or effect size of regressor i. Figure 2.2 gives an example. Its design matrix X has three regressors of interest corresponding to three conditions presented during scanning, and seven nuisance regressors, to account for confounds such as head motion and signal drift.

β is estimated beginning with a randomly initialized estimate $\hat{\beta}$, and then we calculate the estimated signal $\hat{Y} = X\hat{\beta}$. The role of ϵ as the residual error now becomes clear as we have $\epsilon = Y - \hat{Y}$. Least squares regression on $\hat{\beta}$ to minimize $S = \epsilon^T\epsilon$ yields the estimated model. As was first shown by Roger Penrose in 1956[81], the $\hat{beta}$ values which minimize S (with an important caveat discussed below) are obtained by the following "normal" equation:

$$\hat{\beta} = (X^T X)^{-1} X^T Y. \tag{2.2}$$

A **contrast** in the GLM is defined by a set of weights, one for each β_i, which are used to specify a linear combination of the parameters. For example, if there were

three parameters (and thus three regressors) we would have:

$$c_1 * \hat{\beta}_1 + c_2 * \hat{\beta}_2 + c_3 * \hat{\beta}_3 = c^T \hat{\beta}. \tag{2.3}$$

Now say regressor 1 corresponds to the timesteps of the scan where music was played for the participant, and our goal is to find voxels that are active while listening to music. In other words, we want voxels where the GLM has an effect size for regressor 1 (β_1) that is greater than zero with statistical significance. Then we could write our null hypothesis as $c^T \hat{\beta} = 0$ with $c_1 = 1$, $c_2 = 0$, and $c_3 = 0$. Say regressor 2 corresponds to watching video during the scan and regressor 3 corresponds to looking at a still image during the scan. Now if our goal were to find a difference between processing video and images, we would set $c_1 = 0$, $c_2 = 1$, $c_3 = -1$ to obtain a null hypothesis of $c^T \hat{\beta} = \beta_2 - \beta_3 = 0$. We would then look for voxels where we are able to reject the null hypothesis. This is the general strategy for inference with GLM. For a question of interest, we choose values for the c_i to obtain a null hypothesis of $c^T \hat{beta} = 0$. Then the t-statistic at each voxel can be computed as

$$t = \frac{c^T \hat{\beta}}{\sqrt{\text{var}(\epsilon) c^T (X^T X)^{-1} c}}. \tag{2.4}$$

Consider the need to account for the enormous number of statistical tests used in this mass univariate approach. Even low-resolution fMRI scans will have over one hundred thousand voxels, and a collection of this many statistical tests will be rife with false positives. Further, because clusters of voxels may not be spatially independent, simple multiple-comparison methods to correct p-values may be inappropriate. Without burying this work in the details, we note that Gaussian fields are stochastic processes that conform very nicely to realizations of brain scans under normal situations. Within a few years of the advent of fMRI, methods to compute corrected p-values were deeply

embedded in Gaussian field theory[122].

2.2.2. Historical Overview

The GLM itself has a broader history in statistics and regression analysis which predates its application to fMRI data, but we begin our overview with fMRI related work. One of the earliest pioneering studies which brought GLM to fMRI data was published in 1990 by Ogawa et al.[74]. This seminal work introduced the BOLD signal as a correlate of neural activity by using the GLM to model and analyze the signal changes associated with visual stimulation.

The existence of the relationship between activity and measured BOLD signal quickly became well-known. It remained to establish a formal framework for the hypothesis testing of voxel data. In 1994, one of the most cited works in this domain was published, in which Friston et al.[25] presented a simple and complete approach to the hypothesis testing of fMRI data by unifying GLM and Gaussian fields. This approach exists under the umbrella of "statistical parametric mapping" (SPM), which refers to the statistical processes used to test hypotheses about functional imaging data in general, not just fMRI. At this point, SPM methods had yet to address the temporal dependence of successive fMRI images as SPM had hitherto been applied only to PET imaging, which does not have this dependency. To that end, in 1995, Friston et al.[26] published a modification to the SPM approach which used hand-crafted heuristics to allow correlations between the error terms of each timestep. Later that year, in the aptly named "Analysis of fMRI Time-Series Revisited–Again," Worsley and Friston[123] were able to solve the same problem without relying on heuristic arguments. The list of fMRI results published using the above approach is insurmountable, but we present some interesting works here.

Some of the first work in the auditory domain was published in 1997 by Zatorre[127], which employs the GLM to explore the structural correlates of phonetic perception,

melodic processing, auditory working memory, and auditory imagery. In 1998, Buchel et al.[8] used SPM to identify brain regions involved in visual attention modulation. To do this, the authors modeled the four different conditions "attention," "no attention," "fixation," and "stationary" as regressors in their GLM.

One year prior, in 1997, Dale and Buckner[15] employed the GLM to average fMRI data from individual trials, enabling the investigation of rapid cognitive processes in the brain. All of the fMRI datasets in our work use a "block design," which alternates between periods of a stimulus and rest, with the rest period allowing the stimulus-evoked response to return to baseline. However, the ability to analyze closely spaced single-trial signals, or "event-related design", as in Dale and Buckner, still provides for a useful class of investigations[94]. An example that includes the GLM is Ollinger, Shulman, and Corbetta in 2001[75]. The authors used the GLM to separate overlapping sensory, cognitive, and motor components within individual rapid trials by modeling each component with its own regressor.

Adams and Janata(2002)[1] used SPM to identify regions where BOLD signal was modulated during auditory and visual object categorization. More recently, in 2009, Janata[44] identified regions of the brain associated with familiarity, autobiographical salience, and positive emotional affect (valence) (Figure 2.3). They used three sets of regressors that represented 1) a 30s period during which music was playing 2) a question and answer period following each musical excerpt; and 3) a set of parametric regressors that captured the variation in the familiarity, autobiographical salience, and degree of positive affect evoked by each musical excerpt. Their results contributed evidence toward their hypothesis that memories and music are associated in the medial prefrontal cortex.

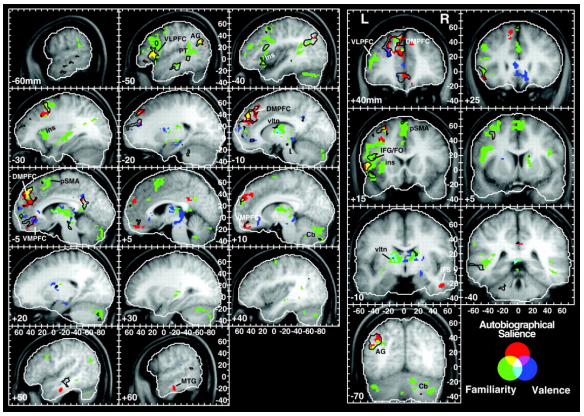

Figure 2.3: Image from [44]. Regions of effects due to music: familiarity (green), autobiographical salience (red), positive emotional affect (valence) (blue), and combined effect of all three (outlined in black).

2.2.3. Conclusions

The GLM/SPM approach has continued from this time to today, but we have presented the landscape leading up to and extending slightly past a new approach to brain decoding introduced in 2001, which is detailed in the next section. To motivate this new approach, note there are two major assumptions inherent to basic GLM analysis, either one of which may be rightfully challenged. First, GLM is a "mass-univariate" approach, calculating statistics one voxel at a time and assuming that signals from each voxel are independent of one another. It may well be, though, that the covariance across neighboring voxels is in fact informative about the cognitive function under examination. Second, the solution for $\hat{beta}$ in Equation 2.2 can only be derived after assuming that the errors ϵ_i are independent and drawn from a normal distribution with mean zero. This assumption may not hold for a variety of reasons, including overlap in the regressors. Indeed, univariate approaches to fMRI data such as contrast subtraction can be useful for basic analysis, but such approaches struggle to isolate the densely overlapping patterns of multivariate signals which comprise neural activity [80, 121].

Section 2.3

MVPA

Multivariate pattern analysis (**MVPA**, sometimes also called multi*voxel* pattern analysis) is a set of methods for analyzing neural responses as patterns of activity. This framing affords investigation of the varying brain states that an area of the brain can produce. This stands in contrast to the GLM which indicates only the extent to which an area of the brain is globally engaged. Thus MVPA methods increase the amount of information that can be decoded from brain activity.

2.3.1. Techniques

MVPA involves searching for reproducible spatial patterns of activity that differentiate across experimental conditions. MVPA is therefore considered as a supervised classification problem where a classifier attempts to capture the relationships between spatial patterns of fMRI activity and experimental conditions.

More generally, classification consists in determining a decision function f that takes the values of various "features" in a data sample x and predicts the class of x. We use "Features" here in the familiar machine learning sense to mean the set of variables or attributes describing x. As a concrete example, x could be a concatenated sequence of timesteps during which a particular stimulus was presented, and the features may represent the corresponding fMRI signals in a cluster of voxels. The different stimuli presented during the scan would then be the class labels.

To obtain the decision function f, data (i.e., samples and their corresponding class labels) must be split into two sets: "training set" and "test set." The classifier is trained using the training set. Training consists of modeling the relationship between the features and the class label by assigning a weight w to each feature. This weight corresponds to the relative contribution of the feature to successfully classify two or

more classes. When more than two classes are present in the experimental design, the analysis can be transformed into a combination of multiple two-class problems (i.e., each class versus all the others). The classifier is then evaluated with the test set to determine its performance in capturing the relationship between features and classes.

This need for a classifier motivated the adoption of early ML architectures for multivariate fMRI analysis [73, 34], notably support vector machines for brain decoding classification [64, 102, 119, 40]. Support vector machines (SVMs) have become popular as supervised classifiers of fMRI data due to their high performance, their ability to deal with large high-dimensional datasets, and their flexibility in modeling diverse sources of data. We do not provide a rigorous mathematical foundation for SVMs here, but we remind the reader that the objective is to separate the different classes while balancing accuracy against overfitting. Figure 2.4 provides instructive examples.

In 2012, Haxby[34] published a paper recounting the developments and innovations of MVPA over the prior decade. He begins the paper by stating, in no uncertain terms, "Multivariate pattern analysis of fMRI data has proven to be more sensitive and more informative about the functional organization of cortex than univariate analysis with the general linear model." MVPA was slow to be adopted though, as it does not provide simple answers to the kinds of questions people were asking at the time[34]–Where is the speech area, or the motor area? Where is reward processed? and so forth. The slow adoption was also because MVPA addressed questions that people hadn't thought of investigating – quoting Haxby, **"What are the varying brain states in an area and how do they encode different types of information?"** This question mirrors quite nicely our rephrasing of Aristotle above, and we will refer back to it in later chapters.

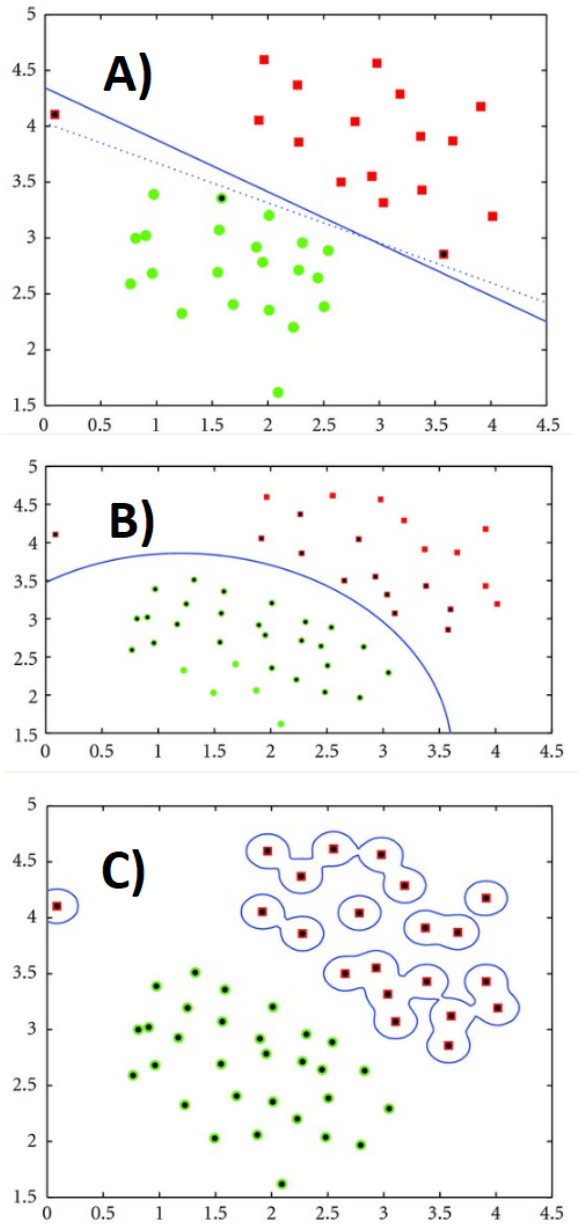

Figure 2.4: Examples of Support Vector Machine (SVM) hyperplanes. A) Linear SVMs allowing some training error (solid) and allowing no training error (dashed). This represents the tradeoff between acceptable classifier performance and overfitting. B) Nonlinear SVM with polynomial kernel $d = 2$, $K = 0$. C) Nonlinear SVM with Radial Basis Function kernel $\sigma = 0.2$. Images from [34].

2.3.2. Historical Overview

Haxby et al. (2001)[35] devised the first prototype MVPA method during their investigation of functional architecture for face and object recognition in ventral temporal cortex. In 2003, Cox and Savoy[12] utilized linear discriminant analysis and support vector machines to classify patterns of fMRI activation evoked by the visual presentation of various categories of objects. Classification was done using only small amounts of data (20s worth) at a time. They achieved classification accuracies well above chance using regions of interest restricted to higher-order object-selective visual areas. Relevant to the transition from mass univariate methods to MVPA, they note: "In contrast to typical fMRI data analysis, in which hours of data across many subjects are averaged to reveal slight differences in activation, the use of pattern recognition methods allows a subtle 10-way discrimination to be performed on an essentially trial-by-trial basis within individuals, demonstrating that fMRI data contain far more information than is typically appreciated."

Results in visual areas of the brain continued. Kamitani and Tong (2005)[46] were able to decode which of eight edge orientations the participant was looking at, as well as which of two overlapping orthogonal gratings they chose to consciously attend on. These results demonstrated that primary visual cortex contains detailed orientation information that can reliably predict subjective perception.

It was not until 2008 that auditory fMRI studies began to use MVPA, when Formisano et al.[24] decoded different vowel sounds and speaker identities from the same set of voxels. Staeren et al. (2009)[103] used MVPA methods to decode four different audio labels: cats, female singers, acoustic guitars, and individual tones. They note they were unable to do so with conventional contrast-based methods. Raizada et al. (2010)[90] performed MVPA analysis on native English and Japanese speakers' neural response to the syllables /ra/ and /la/ in right primary auditory

cortex. They demonstrated that the statistical separability of those responses predicted the participants' behavioral ability to differentiate the syllables. Lee et al. (2011)[53] examined discrimination of melodic contour (the "ups" and "downs" of music) using MVPA. They identified three distinct regions in which the local pattern of activity accurately discriminated between contour categories. Giordano et al.(2011)[27] used MVPA methods to obtain evidence of abstract encoding of non-speech biological sounds, which up to that point had only been known for human speech. Casey et al. (2011) performed 5-way music genre classification using MVPA and an SVM classifier. Deep learning methods began to creep into fMRI analysis around this time and thus we end this subsection here, despite MVPA's continued contributions to the field.

2.3.3. Conclusions

By considering patterns of activity across multiple voxels, MVPA can capture more nuanced and distributed information in the brain, allowing for the detection of subtle differences or patterns that may be missed by univariate methods. On the other hand, the high-dimensional nature of the data and complex models used in MVPA can make it difficult to attribute specific cognitive or neural interpretations to the identified patterns. A further challenge is that MVPA typically requires a relatively large amount of high-quality training data to build accurate models. However, fMRI datasets have continued to grow larger due to higher spatiotemporal resolution from technological advances, and increasingly large sample sizes in general, particularly from big-data initiatives such as the Human Connectome Project[112] and OpenNeuro[84]. Larger datasets can potentially accommodate more sophisticated statistical models than MVPA, with greater power to identify, extract, and distinguish noise sources and signals of interest.[49]

Section 2.4

Deep Learning

We now move on to the discussion of *deep* machine learning techniques which have had empirical success in learning representations of high-dimensional data, without the need for hand crafted features[118] as was the case with MVPA and GLM. Further, the nonlinear activation functions in deep ML models enable the learning of a more complex output function than those that can be learned using traditional machine learning methods. Kuntzelman et al. (2021)[49] give a retrospective view of deep learning with neuroimaging data, in which they suggest that "deep learning has the potential to perform most of the tasks for which traditional MVPA is typically employed, but with greater speed, flexibility, and power," and that deep learning results up to the time of writing "represent only the tip of the proverbial iceberg."

2.4.1. Techniques

Progression into deep ML models has seen multilayer perceptrons[105], autoencoders[79, 43], convolutional neural networks (CNN)[118, 125], and graph neural networks (GNN)[58] for feature extraction and classification of single fMRI images. Architectures designed for time series analysis are desirable due to the high degree of temporal correlation in BOLD responses, and indeed recurrent neural networks (RNN) and various long short-term memory (LSTM) models have been reported[14, 20, 57, 131, 108, 86].

2.4.2. Historical Overview

Since the neural network revolution (or perhaps renaissance) began in 1998 with the Convolutional Neural Network (CNN)[51], there has been consistent and growing interest in applying various deep learning techniques to fMRI data[82]. CNNs, for example, have been successfully applied to the learning of representational features

from fMRI data. In 2014, Yamins et al.[125] showed that optimizing CNNs *solely* on the task of decoding object labels from visual object recognition fMRI data simultaneously optimized the model's ability to predict responses in visual cortex. The authors thus refer to CNNs as "biologically plausible," a quality lacking in GLM and MVPA.

In 2020, Li et al.[58] used GNNs to decode the labels "Autism Spectrum Disorder" and "Healthy Control" from the Autism Spectrum Database[115] as well as the 7 different task labels in the Human Connectome Project (HCP)[112]. In both cases, their models significantly outperformed MVPA-SVM classifiers. Note that this architecture is only capable of classifying one image at a time, as are basic CNNs. The temporal dependence of fMRI data, however, suggests that we look toward architectures which accept sequential inputs.

Dakka et al. (2017)[14] used LSTMs to decode whether the participant had been diagnosed with schizophrenia from sequences of fMRI data evoked by a certain auditory stimulus. The LSTMs outperformed their baseline MVPA-SVM classifiers. Huang et al. (2017)[43] proposed an architecture based on the sparse convolutional autoencoder to learn high-level features from handcrafted time series derived from the raw fMRI data. Wang et al. (2018)[118] proposed a 4-layer CNN that classifies tasks from the raw fMRI voxel values. Their method achieved an average accuracy of 89.0% and 94.7% on a working memory task and a motor classification task, respectively, higher than the accuracy of 69.2% and 68.6% obtained by the SVM-MVPA. A network visualization analysis showed that the CNN automatically detected features from areas of the brain related to each task. Their work used HCP, treating the entire fMRI timeseries as input.

In our previous work in 2021, we trained a sparse autoencoder to reconstruct sequences of fMRI data[79, 64] and used the trained encoder weights as filters in a CNN to transform the labelled fMRI data. The transformed data enabled higher

accuracies on brain decoding tasks with MVPA-SVM classifiers, as well as the learning of a novel decoding task that the SVMs could not learn on either the transformed or original data. We hypothesize that the autoencoder was learning a basis for the vector space of BOLD activity, and thus the filtering process expressed the transformed data as a linear combination of the basis. This compact representation would then be responsible for the improved SVM performance. Note that the autoencoder was trained on the *unlabelled* fMRI data, that is, the images taken in between stimulus trials. Unlabelled data often comprises large amounts of a dataset–in this case it was more than half– but has little to no use in SPM and MVPA strategies. Indeed, the ability to learn relevant latent patterns in unlabelled data is a critical advantage of deep learning over prior techniques. During this work, we were not able to decode the instrumental timbre labels of "Clarinet" and "Trumpet" from either the original or transformed data. BEAT, however, succeeds at this task, as we will discuss in Chapter 5. Recall the description in [34] of MVPA as a new way of asking questions to outperform GLM. We claim that BEAT presents yet another new style of question to empirically outperform MVPA and other deep learning methods when decoding task information from fMRI data.

2.4.3. Conclusions

Deep learning models have enabled performance leaps in high-dimensional fMRI data[7]. However, the application of deep learning models to neuroimaging data poses several challenges, due to the high dimensionality, low sample size, and complex temporo-spatial dependency structure of these datasets. Furthermore, trained deep models generally act as a black-box, which is to say that it is unclear how all the individual parameters work together or on their own to reach a decision. This lack of interpretability impedes insight into the association of cognitive state and brain activity[108]. While deep learning methods are effective if enough data are available

for training, most typical neuroimaging studies have collected data from only tens to hundreds of subjects, with the purpose of identifying minor differences between different states[41] or groups thereof[116]. As noted above, though, deep learning methods can leverage latent patterns in unlabelled data, while MVPA methods alone cannot. Recall from Chapter 1 that transfer learning can help overcome these challenges. While transfer learning strategies do exist for architectures such as RNN and CNN, a different architecture has come to dominate most, if not all, modes of transfer learning with high-dimensional timeseries data–the Transformer. This architecture is the focus of the first section of Chapter 4, and provides the foundation for BEAT. For now, though, we proceed into Chapter 3, which provides a more detailed explanation of fMRI data collection which is necessary to understand the design and implementation of BEAT.

Chapter 3

FMRI Overview and Collection

Functional Magnetic Resonance Imaging (**fMRI**) is an imaging modality for the functional activity of the brain including physical activity, active thought, and response to stimuli. This chapter explains the concepts and background of fMRI studies necessary to understand this thesis and its motivations, while contextualizing those concepts through our data collection in early 2023. **This dataset is one of the main contributions of this thesis**. Section 3.1 introduces the Blood Oxygenation Level Dependent (**BOLD**) signal measured by the fMRI scanner. Section 3.2 explains the motivation, design, and implementation of the scanning protocol for this collection. Section 3.3 explains the extensive preprocessing techniques performed on the data after scanning. Section 3.4 details the previously-existing datasets explored in this thesis. Note that our experiments on the "other datasets" took place before the collection we describe below. The datasets are presented in reverse order to more easily present the necessary concepts in the context of our own work. Finally Section 3.5 details the extraction of specific regions of brain used in our experiments.

Section 3.1

MRI and BOLD Signal

Oxygen is delivered to neurons by hemoglobin in capillary red blood cells. When neuronal activity increases there is an increased demand for oxygen and the local response is an increase in blood flow to regions of increased neural activity (Figure 3.1).

Hemoglobin is diamagnetic when oxygenated but paramagnetic when deoxygenated. This difference in magnetic properties leads to small differences in the reaction of blood to a strong external magnetic field depending on the degree of oxygenation. Since blood oxygenation varies according to the levels of neural activity these differences can be used to detect brain activity. We note that quite a bit of work has been abstracted out into the phrase "detect brain activity," but this is enough to proceed with the relevant discussion. This form of magnetic resonance imaging (MRI) is known as blood oxygenation level dependent (BOLD) imaging. At the beginning of a scan, a high resolution image is taken in order to identify anatomical structures. We refer to this as the "T1 data." The "runs" of a scan begin after collecting the T1 data, during which images are collected faster but at a lower resolution and stimuli are presented or tasks are performed. We refer to this as the "functional data."

Section 3.2

Enculturation Dataset

3.2.1. Motivation

A basic function of cognition is to detect regularities in sensory input to facilitate the unconscious prediction and recognition of future events[2]. During the past 20 years, these predictions have been shown to account for multiple facets of music cognition, including memory[2], emotions[97], pleasure[30], and reward[10]. Multiple studies have

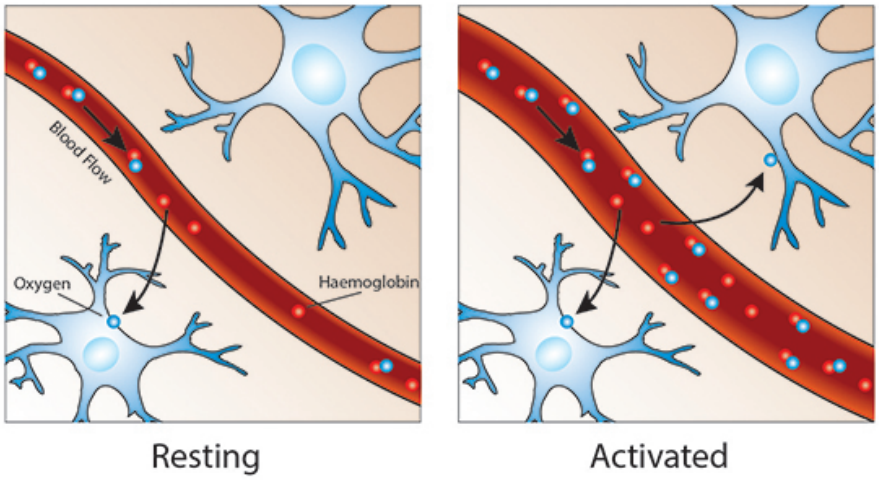

Figure 3.1: (*Top*) Increased blood flow and therefore increased oxygenation consequent to the demand of increased neural activity.

demonstrated behaviorally that listeners' internal models which output the predictions vary across cultures and were consistent with the statistics of the musical grammars of their own musical culture[16, 47, 85]. Studies have also shown that exposure to unfamiliar music induces statistical learning consistent with the exposed music[33, 60]. This mechanism is known as **Musical Enculturation**.

Our goal for data collection was to complement the musical enculturation research conducted by our colleagues Guilhem Marion and Camille Barbarot at l'École normale supérieure (ENS) in Paris. They performed enculturation by instructing 19 participants to listen to unfamiliar Chinese music from the region of Shanxi for at least 30 minutes per day for two weeks. They also instructed a control group of 15 additional participants to listen to at least 30 minutes per day of Bach chorals for two weeks. The relevant musical grammars present in the Bach chorals are generally consistent across all western music, but are distinct from the Shanxi music. We omit further music-theoretic details here. Their main hypotheses were: 1) the enculturation group would update their internal models to make better predictions about Shanxi music than the control group, and 2) the self-reported pleasure ratings would increase in the enculturation group

after learning the Shanxi musical grammar, but not in the control group.

Marion and Barbarot used electroencephalography (EEG) to measure the electrical response to Shanxi samples in the brains of both groups before and after the two weeks of exposure, and in both instances asked the participants to rate the pleasure they felt while listening. The literature on the neural underpinnings of musical expectation indicates that the degree of surprise due to a note is encoded in the measured electrical activity around 200ms from the note onset, with greater amplitude for unexpected notes[18, 52, 63, 76]. In other words, the amplitude around 200ms after note onset serves as a measure of surprise. This approach was used to evaluate hypothesis 1), and the changes in response to Shanxi music induced by the two weeks of exposure – that is, response after exposure minus response prior – are given in Figure 3.2. Both lines peak around 200ms. This appears to indicate that the change in surprise is *positive*, that is, the surprise has gone *up*, for both groups, when it was expected to remain constant in the control group and go down in the enculturation group. However, this effect is explained by an increased signal-to-noise ratio (SNR). Indeed, participants' familiarity with an experiment generally improves SNR along with generating larger responses. The important conclusion, rather, is that the surprise experienced by the enculturation group is significantly lower than the control group after exposure. This result is consistent with the literature and supports hypothesis 1). Figure 3.3 is more straightforward. It depicts the change in pleasure ratings for the two groups, in which we observe an increase in pleasure ratings for the enculturation group and a decrease for the control group. This is consistent with the literature and supports hypothesis 2).

Now consider that the literature on musical reward suggests that the nucleus accumbens (**NAcc**) is highly involved in behavioral musical pleasure, musical predictions, and musical learning[10, 128]. In particular, surprise and uncertainty from

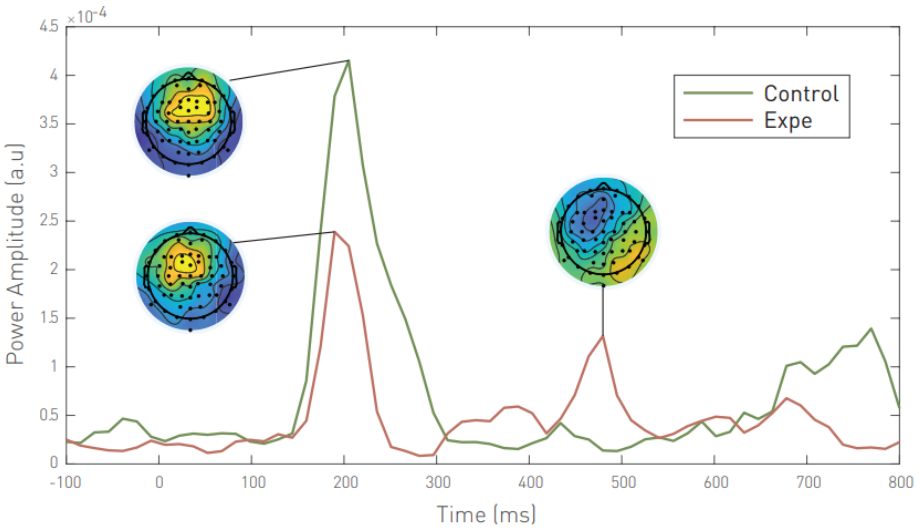

Figure 3.2: The change in response to Shanxi music (after minus before) as measured by EEG. Greater amplitudes around 200ms after note onset correspond to greater surprise, that is, a more severe violation of the internal model's prediction for that note. Both lines are (unintuitively) positive around 200ms, but this is due to a higher SNR after the exposure period. Given that there was no difference in surprise between groups before the exposure period, we conclude a significantly lower degree of surprise experienced by the enculturation group after exposure.

statistical models of music jointly predict self-reported musical pleasure and evoked activity in NAcc[10]. Thus we expect to see increased activity in both NAcc and STG after enculturation. However, the spatial resolution of EEG does not permit this analysis. Indeed, the total spatial information granted by the EEG study can be seen in the topographical maps in Figure 3.2. Therefore, we sought to complement the work of Marion and Barbarot by measuring the effects of musical enculturation via the same Bach and Shanxi music clips with fMRI rather than EEG. The high spatial resolution of fMRI would then permit the direct analysis of NAcc.

Several other regions of the brain are implicated in musical enculturation. Nan et al. (2008)[68] demonstrated increased BOLD signal in response to culturally unfamiliar musical grammars in right angular gyrus and middle frontal gyrus (possibly reflecting higher demands on attention systems), and the right posterior insula (suggesting higher loads on basic auditory processing). In response to culturally familiar music,

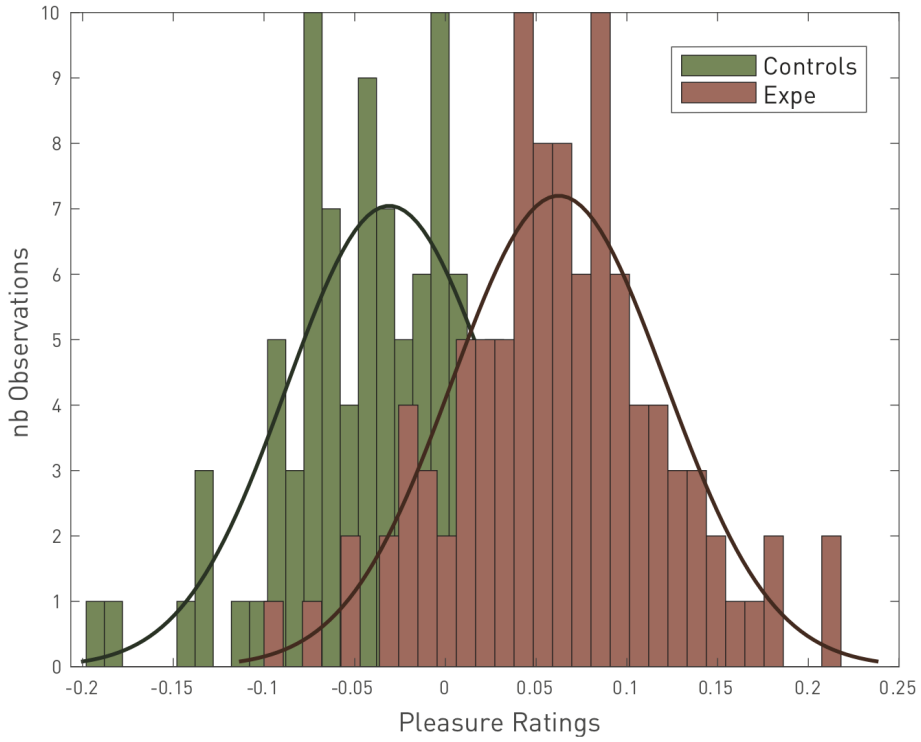

Figure 3.3: Change in pleasure ratings (after - before). An increase (mean above 0) means that participants increased their liking of the pieces.

Nan et al. observed increased BOLD signal in left planum temporale, right ventromedial prefrontal cortex, and bilateral motor regions, the last of which is likely due to "grooving" to the more familiar music, even as one tries to remain still in the scanner. The dorsal striatum as well is implicated in music memory [44, 106], and thus likely contributes to the internal prediction model. In this thesis we consider only Nucleus Accumbens due to time constraint and our motivation to supplement the work of our colleagues Marion and Barbarot.

3.2.2. Methods

The high-level design for this data collection was as follows: 1) Recruit participants familiar with western musical grammars and unfamiliar with Shanxi musical grammars; 2) Expose participants to clips of Bach and Shanxi during fMRI scan and periodically ask them to rate the pleasure of the music; 3) Expose participants to at least 30

minutes of Shanxi music at home every day for one week following the initial scan; 4) Expose participants to Bach and Shanxi during a second fMRI scan at the end of the exposure week; 5) Attempt to train a classifier to distinguish the two scanning sessions. We now explain each of these steps in detail.

1) Recruitment of participants willing and available to commit to two scans one week apart is difficult. Nevertheless we recruited 5 participants. Each was at least 18 years old and gave their written informed consent for each scan in accordance with the Institutional Review Board at Dartmouth College. They completed a brief questionnaire to determine eligibility in which four participants responded that they listened to at least 5 hours of music per week and the fifth between 0 and 5 hours. All participants responded that they had actively listened to western classical music during more than 5 years of their life, and Chinese folk music during 0 years. All five were thus deemed eligible. Upon arrival for each scan, the participants filled out a screening form to confirm they could be scanned safely. They were each compensated $60 USD after the second session.

2) Each scan consisted of 8 runs. Each run began with two dummy TRs and then consisted of four "blocks," which themselves consisted of four "trials." The design of a single trial is shown in Figure 3.4. All trials in a given block are the same style, resulting in 48 trials for each style per scan. There is no time between trials. A randomized jitter value between 4 and 7.5 seconds is assigned to the beginning of each trial to decouple the evoked response from elapsed time and prevent a consistent expectation of music starting. The scanning parameters were as close to those of the Genre Dataset as possible to facilitate transfer learning from one to the other. In particular, $1mm^3$ voxels and a 1.5s TR.

The music clips were procedurally generated by Marion and Barbarot using a synthesizer according to the rules of the two musical grammars. Both styles of clips

are simply a sequence of notes played on the Guzheng, a Chinese instrument. This construction controls for effects due to different instruments and complex melodic interactions, thereby focusing the experiment on the differences in the musical grammars.

The EEG experiment by Marion and Barbarot asked participants to rate the pleasure experienced by each clip on a continuous scale from 0 to 1, and they report a significant increase in the pleasure ratings after exposure to Shanxi music. However, in our fMRI scanner, the only method of feedback from the participant is a controller with four buttons. Thus we prompted the participants to rate the pleasure they experienced as one of "Not pleasing," "Somewhat pleasing," "Moderately pleasing," or "Very pleasing." The lack of complexity in the music combined with the relative weakness of our rating system makes this a difficult and imprecise question, and thus we did not expect to reproduce the increased pleasure results of Marion and Barbarot. This prompt still serves to monitor the attention of the participant. On the other hand, we felt that presenting this prompt after every trial, that is, every 45 seconds, would become tedious and perhaps annoying for the participant, which could corrupt our measurements of NAcc. Thus the prompt was presented after every three trials, that is, at the end of each block.

3) Marion and Barbarot created a website that played Shanxi clips (distinct from clips used during scanning) and recorded the time spent listening. The participants were instructed to listen to at least 30 minutes per day, with 1 hour being preferable. In order to maintain the attention of the participant, the music stopped at random times and required the participant to manually resume it. The listening times for all participants are given in Table 3.1. Day 0 is the day of the first session, and the final listed row is the day before their second session. Two participants had an additional day of listening before their second session due to scheduling issues, but

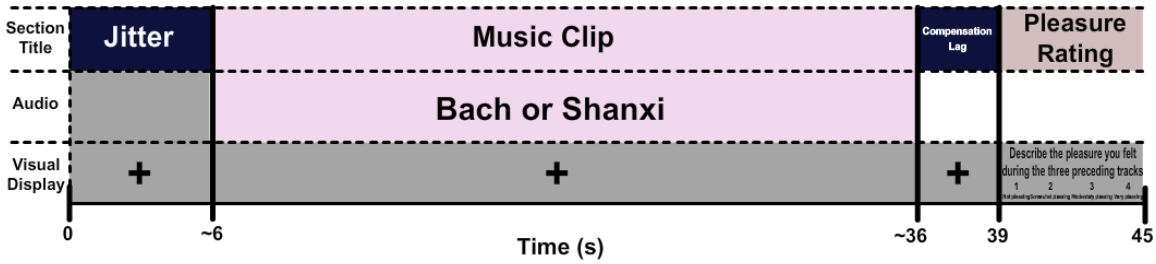

Figure 3.4: The design of each trial during scanning for the Enculturation Dataset. A randomized jitter value between 4 and 7.5 seconds is assigned to the beginning of each trial to decouple the evoked response from elapsed time and prevent a consistent expectation of music starting. The compensation lag is calculated such that the Pleasure Rating prompt appears after 39s, although this prompt only appears at the end of each block. Each participant's functional data consists of 8 runs, each of which had 4 blocks with 3 trials in each block. Each block was either all Bach or all Shanxi. Half of all blocks for each participant were Bach and the other half Shanxi. The arrangement of blocks was randomized for each participant. The two sessions for each participant had identical stimuli presentation.

Table 3.1: Time spent listening to Shanxi music clips during the week of exposure after the first scan, in minutes and seconds. Columns are for the day number and each participant. The clips listened to during this period were distinct from those played during scanning. Day 0 is the day of the first scan and does not include the 24 minutes spent listening in the scanner. Each participant's final listed row is the day before their second scan. Two participants had an additional day of listening due to scheduling issues, but in some sense this is balanced by both participants also having at least one day where they did not listen at all.

Day	Part. 1	Part. 2	Part. 3	Part. 4	Part. 5
0	0s	15m	18m06s	0s	0s
1	33m03s	38m57s	11m48s	41m59s	8m13s
2	34m36s	48m22s	83m48s	40m42s	79m56s
3	48m57s	28m47s	8m21s	48m49s	38m35s
4	0s	39m26s	16m35s	60m0s	17m12s
5	31m22s	42m21s	30m0s	51m19s	0s
6	25m37s	90m37s	79m37s	31m29s	11m8s
7	44m23s				63m49s

this is balanced somewhat by the fact that both had at least one day where they did not listen at home.

4) Toward the goal of distinguishing between the two sessions, the sequence of

clips presented during the second session was identical to the first session in order to allow direct comparison of the neural responses. This does raise the concern of effects due to recognizing the music during the second scan. However, as explained in 2) above, the music is very simple with no unique identifiers. After a week of listening to similar clips at home, we contest that the likelihood of recognizing any of the music during the second scan to be negligible.

5) We present the architecture, task, and training data of this classifier in Chapter 4. The general idea though is to separate the data corresponding to Bach and Shanxi, reduce the data to only Nucleus Accumbens, and then train separate classifiers to distinguish between session one and two for each style. The Bach training serves as a control since our experiment design does not attempt to change the participants' response to western music. Specifically, if the models learn to distinguish the sessions equally well on both styles, despite taking no action to modulate the Bach response, then the difference between sessions is most likely due to confounds such as the participants' increased comfort in the scanner during the second session. On the other hand, if the models have a greater ability to distinguish sessions for Shanxi than Bach, we will attempt to draw conclusions about enculturation.

Section 3.3

Data Preprocessing

The BOLD signal measured by fMRI is typically mixed with non-neural sources of variability[88]. Preprocessing identifies the nuisance sources and reduces their effect on the data[59, 9], and further addresses particular imaging artifacts and the anatomical localization of signals[104]. Extracting a signal that is most faithful to the underlying neural activity is crucial to ensure the validity of inference and interpretability of results[4]. Thus, a primary goal of preprocessing is to reduce sources

of false positive errors without inducing excessive false negative errors. An illustration of false positive errors familiar to most researchers is finding activation outside of the brain due to faulty spatial normalization. As a more practical example, Power et al. (2012)[87] demonstrated that unaccounted-for head-motion in resting-state fMRI generated systematic correlations that could be misinterpreted as functional connectivity. Conversely, false negatives can result from a number of preprocessing failures. We turn to a standardized preprocessing pipeline to achieve an empirically sound balance between these two concerns.

3.3.1. fMRIPrep

We use a standardized pre-processing pipeline called fMRIPrep[22] on all datasets in this work. fMRIPrep is designed to provide an easily accessible, state-of-the-art interface that is robust to variations in scan acquisition protocols and that requires minimal user input, while providing easily interpretable and comprehensive error and output reporting.

The descriptions in this subsection are mostly adapted from the fMRIPrep website[22], but the images are from the output of running fMRIPrep on the dataset we collected in the previous section. All of our preprocessing steps were performed on the Brainlife service[36]. Preprocessing fMRI data requires intense storage and computing resources, and Brainlife provides millions of free computing hours supported by NSF and crowdsourced cycles to researchers and students free of charge. Brainlife also provides secure cloud storage for raw data as well as data derivatives.

The first objective of the pipeline is to remove the skull from the images, in other words, we want to obtain a brain mask (Figure 3.5.

Then, spatial normalization to a standard space is performed using mutual-information based, nonlinear registration scheme. We chose the MNI152NLin2009cAsym template. We refer to this as "**MNI space**" throughout. Figure 3.6 shows two dif-

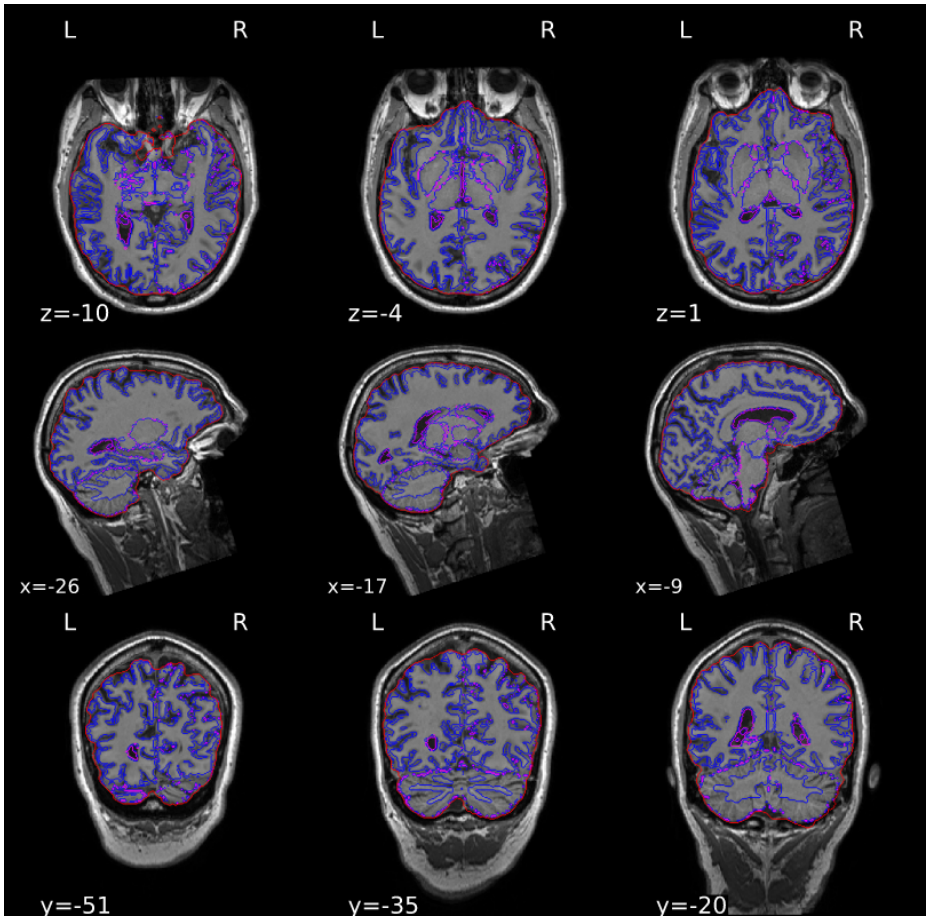

Figure 3.5: fMRIPrep uses an atlas-based method for skull extraction. The output includes a single figure overlaying the brain mask (red), and tissue boundaries (blue = gray/white; magenta = tissue/Cerebro-Spinal Fluid (CSF)).

ferent participants in the first column, and their transformations to MNI space in the third and fourth columns. The circled regions in the figure are easily identifiable dissimilarities between the two participants that have become indistinguishable in MNI space.

Next we need to reconstruct the surface boundaries in MNI space (Figure 3.7). fMRIPrep can outsource this step to another application called Freesurfer[23], which we were able to do on Brainlife.

The preprocessing of the T1 data is complete. It remains to preprocess the functional data. The first step is head-motion correction. To do this, the first image of

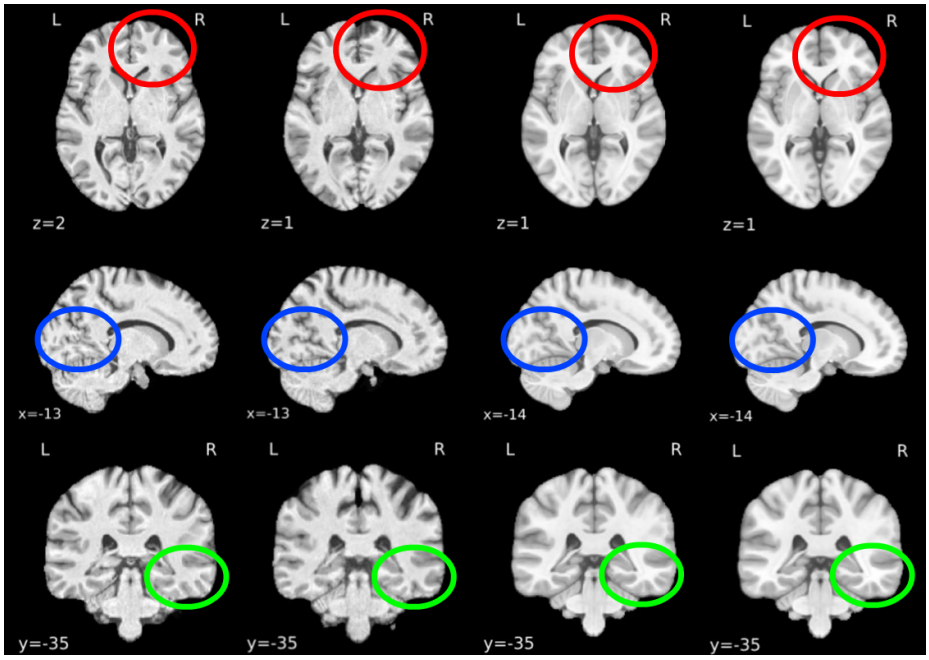

Figure 3.6: Spatial normalization of the T1 image to the MNI152NLin2009cAsym template. Columns from left to right: Participant A in T1 space, Participant B in T1 space, Participant A in MNI space, Participant B in MNI space. The circled regions are easily identifiable dissimilarities between participants that have become nearly indistinguishable in MNI space.

each run is chosen as the reference image. Then a rigid-body transform is calculated for each timestep with respect to the reference image, and then applied to that timestep to correct for head-motion. The second step is slice-time correction. When we perform analysis of fMRI data, we treat each TR as if the entire image were obtained instantaneously. The reality is that each image is collected in slices over the course of the TR. For example, suppose an image consisted of two slices, and slice 2 were acquired 0.1 seconds after slice 1. Then either slice 2 would need to be shifted (interpolated) back in time 0.1 seconds, or slice 1 would need to be shifted 0.1 seconds forward in time, in order for us to treat the image as an instantaneous snapshot of the brain. This is slice-timing correction, and fMRIprep realigns all slices in time to the middle of each TR.

The next step is to transform the functional data to MNI space by aligning it with

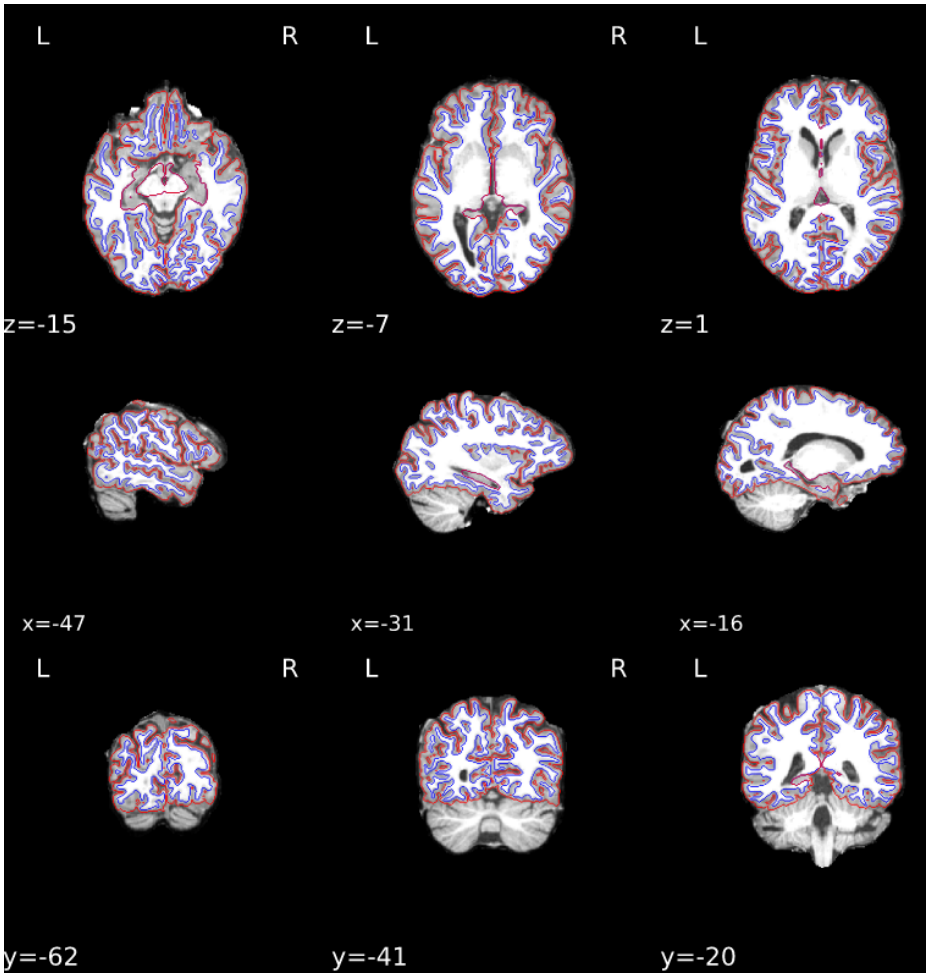

Figure 3.7: fMRIPrep uses the output of Freesurfer to reconstruct the surface boundaries in MNI space.

the T1 image in MNI space we obtained earlier (Figure 3.8).

Finally we calculate the brain mask for the functional data in MNI space (Figure 3.9).

All remaining steps are performed on the output of this preprocessing pipeline.

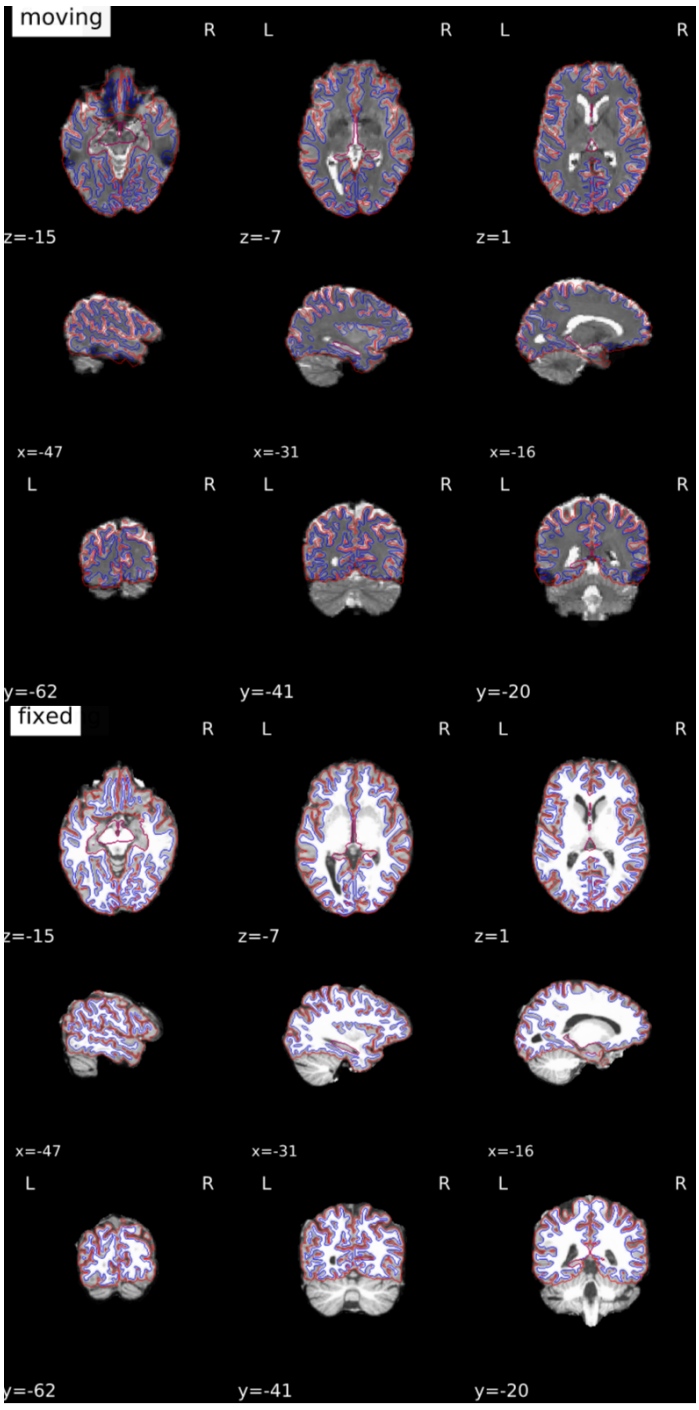

Figure 3.8: Mapping functional data to MNI space by aligning to T1 reference image in MNI space.

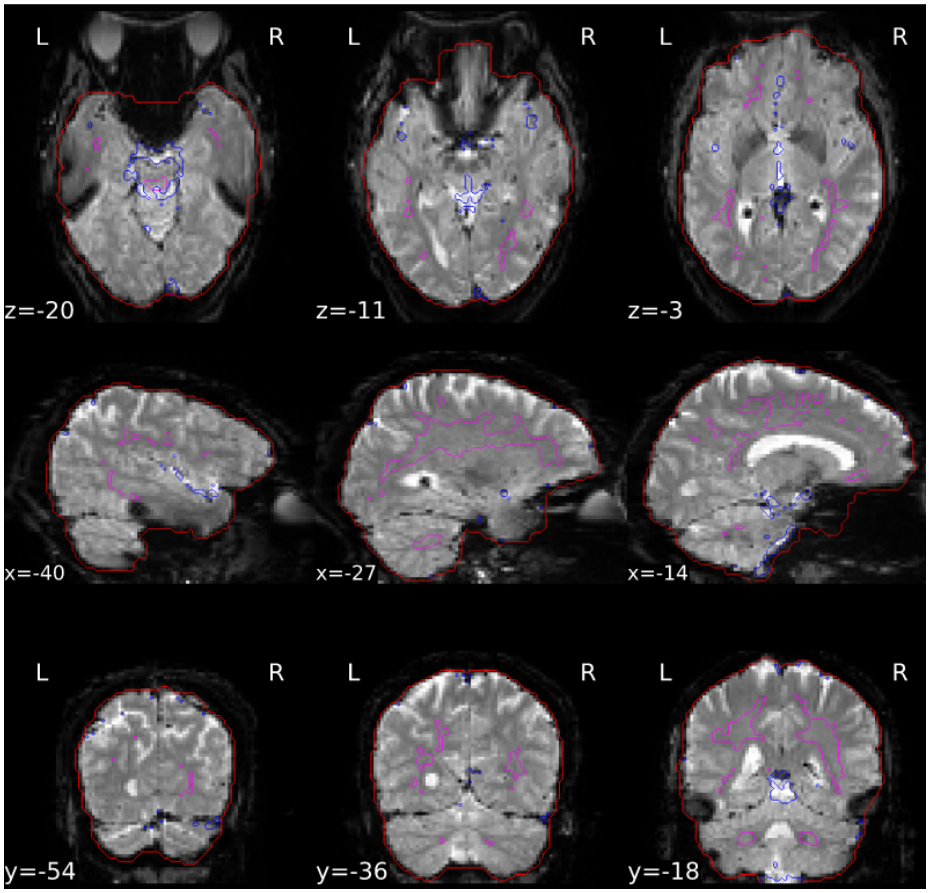

Figure 3.9: Calculating the brain masks for the functional data in MNI space.

Section 3.4

Other Datasets

We collected the Enculturation dataset in early 2023. Until that point our experiments focused on two different datasets. We present their details here as they are critical to understanding the experiments.

3.4.1. Auditory Imagery Dataset

This dataset was collected by May et al. in 2020 and has not yet been made officially public, but is tentatively available upon request. On other hand, work built on this dataset has been published in our lab's previous work. [64, 79].

Candidates to participate possessed at least 8 years of formal music training or professional performance experience in Western tonal music, and they completed the Bucknell Auditory Imagery Scale[31] and the Bregman Musical Ability Rating Survey[32]. Twenty-three such participants passed the screening process and provided their written informed consent in accordance with the Anonymous Review Board. Each subject was compensated $20 US upon completion of the scan.

All scanning used a 3.0 T Siemens MAGNETOM Prisma MRI scanner with a 32-channel head coil and Lumina button box with four colored push buttons. Each functional scan performed a T2* weighted single shot echoplanar (EPI) scanning sequence with a repetition time (TR) of 2 sec and 240mm field of view with 3mm voxels, yielding 80 voxel by 80 voxel images with 35 axial slices for a total of 224,000 voxels per volume. We used the *fmriprep* software[22] to perform motion correction, field un-warping, bias field correction, and mapping to the standardized MNI space, as well as brain extraction and ROI parcellation, on the raw T2* BOLD data. As mentioned the dimensions of the raw functional data were 80x80x34, while the dimensions after being mapped to MNI space were 65x77x65. A sample image from before and after this standardization are shown in Figure 3.10.

Each participant's functional scan consists of 8 runs of 21 musical trials. The design of each trial is depicted in Figure 3.11. Each scan was randomly assigned either the key of E Major or F Major, which was not known by the participant. Each run collected data for either the heard condition or the imagined condition, alternating from run to run. The conditions of the first four runs were repeated in the last four runs. In other words, the task-condition run sequences for each participant were one of either [HT, HC, IT, IC, HT, HC, IT, IC] or [HC, HT, IC, IT, HC, HT, IC, IT].

Each trial began with an arpeggio in the assigned key for the participant to internally establish a tonal context, followed by a cue-sequence of ascending notes in

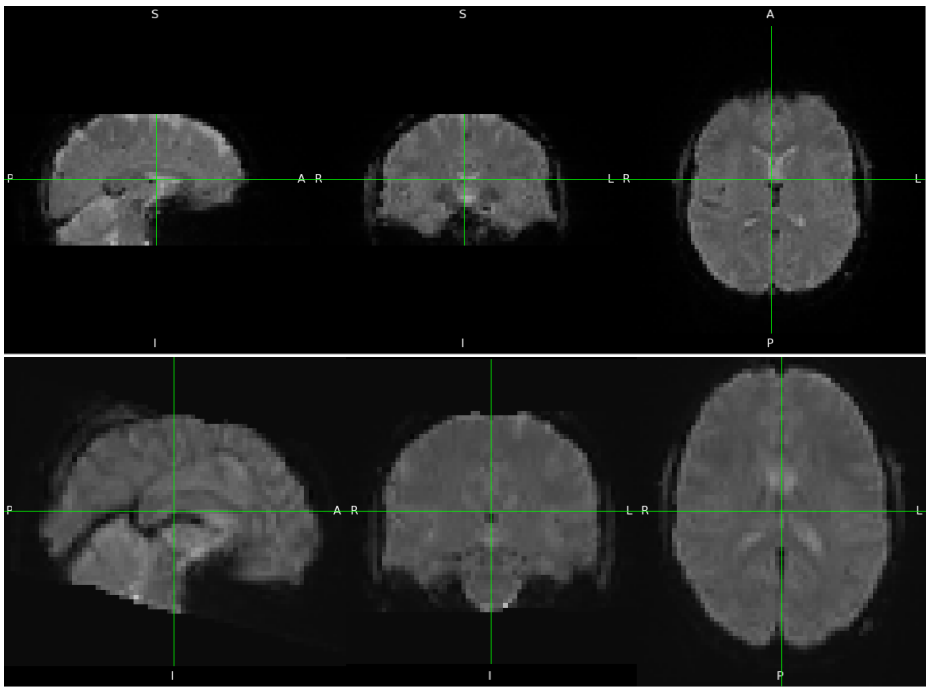

Figure 3.10: (*Top*) Sample raw image from the Auditory Imagery dataset. (*Bottom*) The same image as above after being mapped to the standardized MNI space.

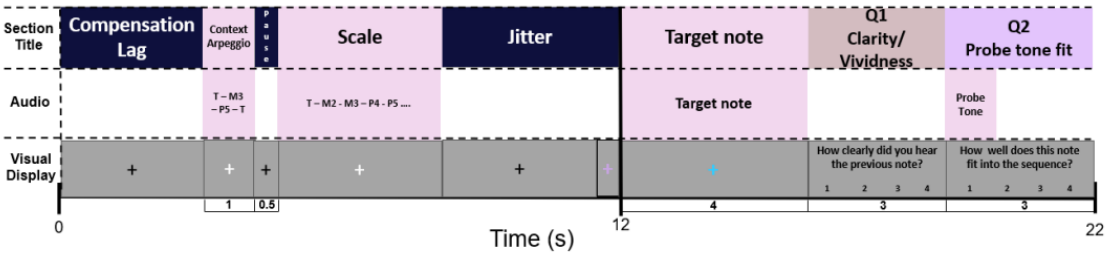

Figure 3.11: The design of each trial during scanning for the Auditory Imagery Dataset. Each participant's functional data consists of 8 runs, each of which had 21 trials.

their assigned major scale. After a randomized time interval, the participant either heard the next ascending note in the scale, or was instructed to imagine the next ascending note, depending on the run. The following four seconds (2 TRs) of scanning collected from all trials constituted the labelled data for the heard and imagined tasks. Next, a probe tone was played, and the participant rated the probe tone's goodness of fit in the tonal context from 1 to 4. They excluded the data of any participant with at least 20% of their ratings missing, or whose ratings did not reflect adequate

understanding of the probe tones' goodness of it. These exclusions controlled for insufficient musical knowledge as well as inattention. Thus they excluded the data of six of the twenty-three participants.

3.4.2. Music Genre Dataset

This dataset was published by Nakai et al. in 2017[67]. Scanning was performed using a 3.0 T MRI scanner equipped with a 32-channel headcoil. For functional scanning, they scanned 68 interleaved axial slices with a thickness of 2.0 mm without a gap using a T2*-weighted gradient echo multi-band echo-planar imaging sequence (repetition time (TR) = 1.500 ms, echo time (TE) = 30 ms, flip angle (FA) = 62°, field of view (FOV) = 192×192 mm^2, voxel size = 2×2×2 mm^3, multi-band factor = 4). For anatomical reference, they acquired high-resolution T1-weighted images of the whole brain from all participants using a magnetization prepared rapid acquisition gradient echo sequence (MPRAGE, TR = 2.530 ms, TE = 3.26 ms, FA = 9°, FOV = 256×256 mm^2, voxel size = 1×1×1 mm^3).

Music stimuli from 10 genres (blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock) were taken at random from the GTZAN music genre dataset[110]. A total of 54 music pieces (30s, 22,050 Hz) were selected from each genre, providing 540 music pieces. A 15-s music clip was selected at random from each music piece. They scanned each participant for 18 runs of 10 minutes each: 12 were considered as "training" runs, and 6 were considered as "test" runs. We emphasize here that the words "training" and "test" in the original run labels have no relation to our own training and validation splits. Each "Training" Run corresponds to 40 different music clips with no break in between clips ($40 \cdot 15s = 600s$), while each "Test" Run corresponds to a sequence of 10 music clips (one from each genre) repeated four times with no breaks ($10 \cdot 4 \cdot 15s = 600s$). In our work, we only considered the *first* instance of each clip in the "Test Runs," to avoid any effects from repeat exposure, meaning

each "Test" Run contributes 10 clips. Thus each of the five subjects contributes $12 \cdot 40 + 6 \cdot 10 = 540$ clips.

The dimensions of the raw functional data were 96x96x68, while the dimensions after being mapped to MNI space were 97x115x97. We wanted to match the dimensions of the audimg dataset to facilitate ease of use of our models and directly compare performance. Thus the MNI data was downsampled via linear interpolation to 65x77x65. A sample image in each of the three dimensionalities is shown in Figure 3.12.

Section 3.5

Regions of Interest

The full MNI space is several orders of magnitude too large for our purposes, and more importantly we are only interested in regions of interest (ROIs) of the brain that may be related to the musical information we are interested in. In this section we first cover the ROI extraction for the Music Genre and Auditory Imagery experiments, and then the ROI extraction for our Enculturation Dataset experiments, as those decisions were informed by our experience with the other two datasets.

3.5.1. Genre and Auditory Imagery

The Superior Temporal Gyrus (STG) is the site of the auditory cortex, which processes auditory information. Angulo-Perkins et al. (2014)[3] showed preferential involvement of STG in processing music in both musicians and non-musicians, which fits our goal of learning from the Music Genre dataset. STG has also been used to learn decoding models of complex natural sounds[32], language[103], and even imagined sound[64, 79] from fMRI data.

More specifically to the Music Genre Dataset, Nakai et al. (2021)[67], the white paper for the Music Genre dataset, indicated distinct cortical organizations for different

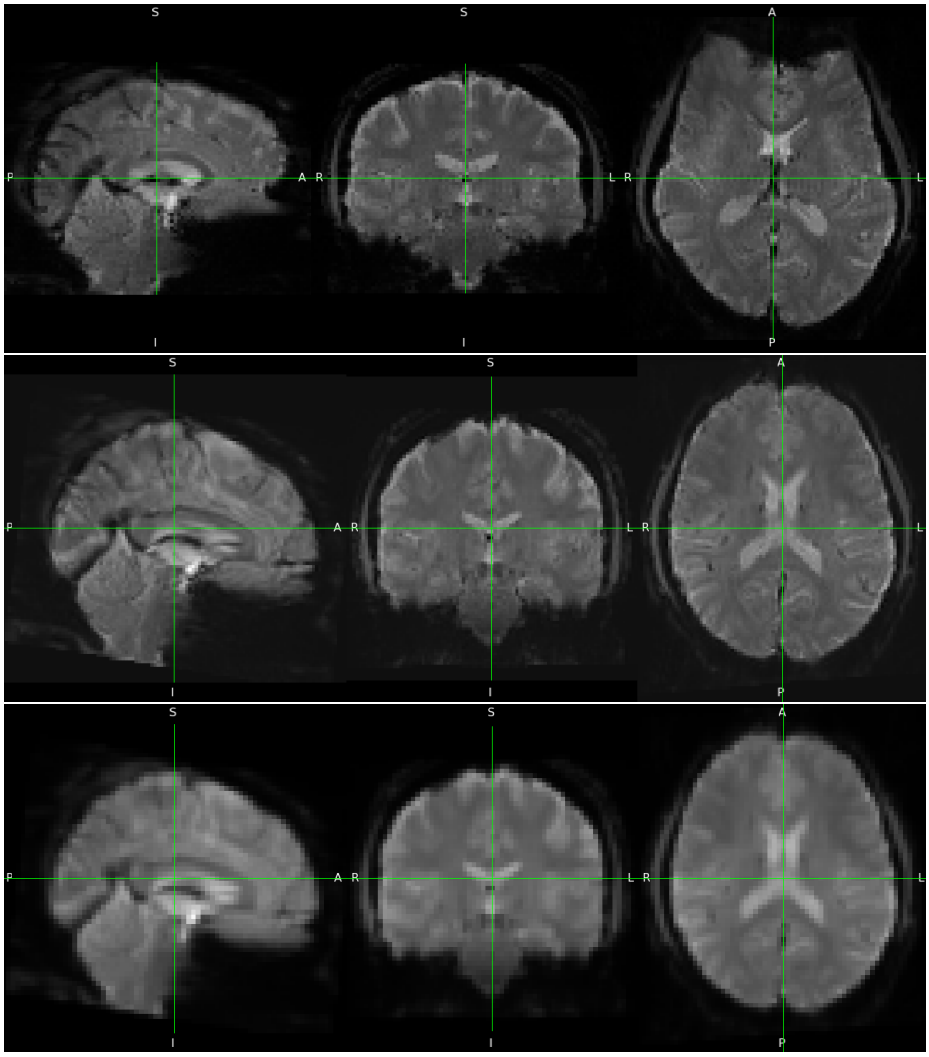

Figure 3.12: (*Top*) Sample raw image from the Music Genre dataset. (*Middle*) The same image as above after being mapped to the standardized MNI space. (*Bottom*) The same image as above after being downsampled via linear interpolation to match the dimensions of the Auditory Imagery data in MNI space.

music genres in the bilateral STG. Specific to the Auditory Imagery Dataset, our previous work successfully decoded heard and imagined pitch from bilateral STG[79, 64]. Thus we focused on STG for both the Music Genre and Auditory Imagery experiments. We proceeded with Left STG and Right STG separately for reduced model complexity and thus lower resource demand for training.

FSLEyes (2022)[65] is a free application for viewing fMRI images and includes

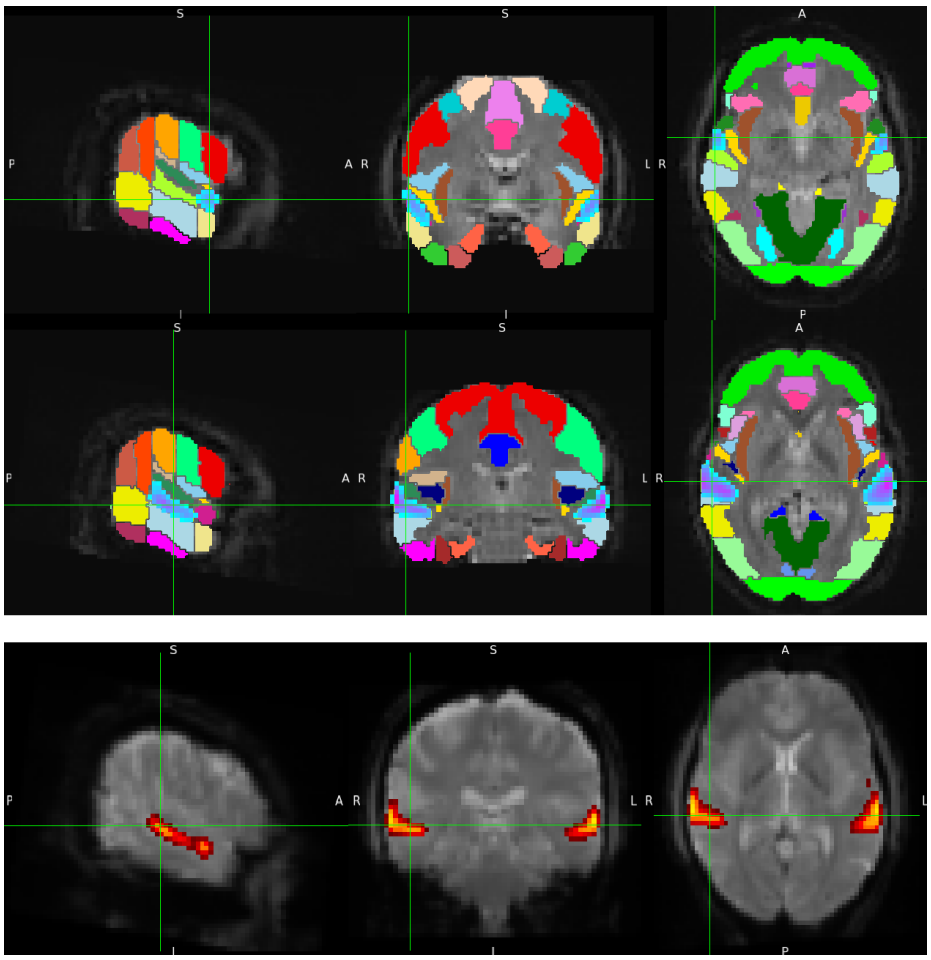

Figure 3.13: (*Top*) The Harvard Oxford Cortical Atlas. (*Bottom*) Heatmap for probability of voxel inclusion in STG. Only probabilities greater than or equal to 23% are shown.

several **atlases** for isolating structural ROIs in the brain with respect to MNI space. We used the Harvard-Oxford cortical structure atlas (HO atlas), some regions of which are shown as an example in Figure 3.13.

The HO atlas assigns a probability to each voxel of belonging to each ROI. Therefore in order to extract STG, we needed to choose a minimum probability threshold for inclusion in STG. This threshold is a hyperparameter to be tuned in future work, but in this work all datasets have a threshold of 23%. We obtained our threshold by visual inspection of the resulting regions. In their seminal work, Craddock et al. (2011)[13]

used a threshold of 25% with the HO atlas, so our visual inspection method was only slightly more lenient.

The HO atlas labels anterior and posterior STG separately, so we applied the threshold to both regions and concatenated them. Voxels which met the threshold for inclusion in *both* anterior and posterior are only included once and the greater of the two probabilities is preserved. Figure 3.13 shows the heatmap corresponding to this union.

The number of voxels in Left STG with inclusion threshold 23% is 413. However, as we will explain in Chapter 3, we wanted the input space to be a round number with a diverse factorization, and either 2 or 3 (depending on the experiment) additional dimensions must be reserved. Thus we chose to insert either 5 or 4, respective to 2 and 3, additional voxels with maximal probabilities below 23%. The number of voxels in Right STG with 23% inclusion threshold is 431, so for this region we removed the 13 or 14 voxels with minimal probability of inclusion, with respect to whether 3 or 2 dimensions were reserved. For both hemispheres this resulted in a 420-dimensional input space for the experiments using inputs constructed from these datasets. We then performed voxel-wise linear de-trending across the full scan and, finally, standardized each voxel to mean zero and standard deviation 1 across the full scan.

3.5.2. Enculturation

Recall from Section 3.2.1 that we expect to see the effects of musical enculturation in several regions related to auditory and emotional processing, and that we focus here on NAcc as a continuation of previous work on the evolution of the internal prediction model. Additional regions are left to future work. The experiments on this dataset (5.3) thus demonstrate BEAT's capacity for transfer learning in regions beyond STG.

We relied on an atlas to extract STG from the previous two datasets, but we take a more deliberate approach with NAcc. One of the outputs of Freesurfer is a

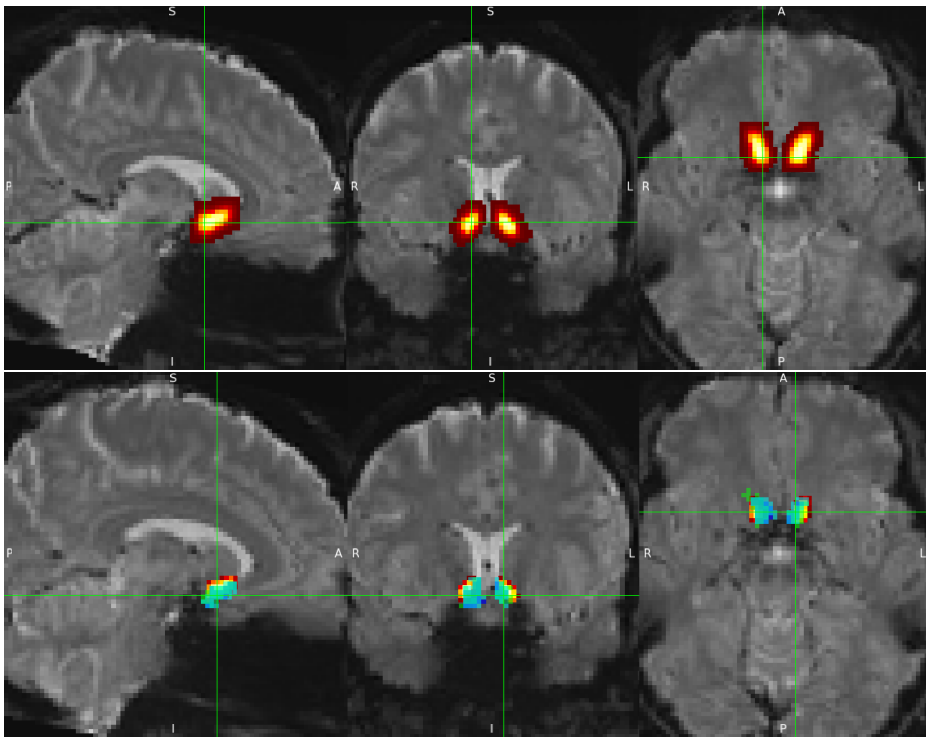

Figure 3.14: (*Top*) Nucleus Accumbens ROI from Harvard-Oxford atlas with threshold 0%. (*Bottom*) Same ROI as top with threshold 23%. Above it are three participants' warped ROIs in transparent green, light blue, and dark blue.

parcellation, which outlines the various structures in the brain. We used an application on Brainlife to obtain a volumetric ROI for Nucleus Accumbens from the outline provided by Freesurfer. These ROIs are specific to each participant's high resolution T1 data. A final Brainlife application warps these ROIs to MNI space. Figure 3.14 compares these warped ROIs to that of the Harvard Oxford Atlas. The top image shows the atlas NAcc with a threshold of 0%. The bottom image shows the atlas ROI as the same heatmap as the top image with threshold of 23%. Three participants' warped ROIs are overlaid in transparent green, light blue, and dark blue. As with STG, a 23% threshold is a satisfying approximation. The union of all participants' warped ROIs yields total 417 voxels. This union is our final Nucleus Accumbens ROI in this work. We performed voxel-wise linear de-trending on the extracted data, and then standardized to mean 0 and standard deviation 1.

Chapter 4

Bidirectional Encoders for Auditory Tasks (BEAT)

The contributions of this thesis include a novel deep learning framework for fMRI-based studies called BEAT, which stands for Bidirectional Encoders for Auditory Tasks. These contributions are the focus of this chapter. In Section 2.4.2 we detailed the history of machine learning with fMRI data, leading to the motivations for the architecture and task designs detailed in this chapter. Section 4.1 explains the Transformer deep learning architecture, on which my framework is based, and presents an overview of how Transformers have been used on fMRI data. In Section 4.2 I present my Transformer-based paired-sequence architecture for fMRI tasks which serves as the model of interest for BEAT. Then in Section 4.3 and Section 4.4 I present the novel training tasks learned by BEAT.

Section 4.1

The Transformer Architecture

The transformer architecture[113], depicted in Figure 4.1, is proposed to supersede Recurrent Neural Networks (RNN)[83], and more specifically, the popular variants of RNN such as Long Short Term Memory (LSTM)[39] and Gated Recurrent Units (GRU)[11], for processing sequential inputs. Most competitive sequence-to-sequence models have an encoder-decoder structure. In the case of the transformer, the encoder (left side of Figure 4.1 maps an input sequence of symbol representations (x_1, ..., x_n) to a sequence of continuous representations z = (z_1, ..., z_n). Given z (the arrow from left half to right half in the figure), the decoder then generates an output sequence y = (y_1, ..., y_m) of symbols one element at a time. At each step the model is auto-regressive, consuming the previously generated symbols as additional input when generating the next. Consider machine translation as an example. The input sequence x is the sequence to be translated, its encoded representation z is passed to the decoder, y is currently an empty sequence, and the decoder outputs the first word y_1 of the translation. Then this repeats with the sequence (y_1) as input to the decoder and y_2 is output, and so on.

For sequence processing, one of the most important challenges is to model the interaction between tokens. In RNNs, input tokens interact with each other through the recurrent function and decide the output for the next recurrent step. The transformer architecture, on the other hand, completely relies on the attention mechanism[6] for token interaction. More specifically, it uses self-attention to model relationships between the elements of the sequence. These relationships are quantified as attention scores, which then becomes weights on each element due to their relevance to other frames. However, without any recurrent or convolution components, the transformer

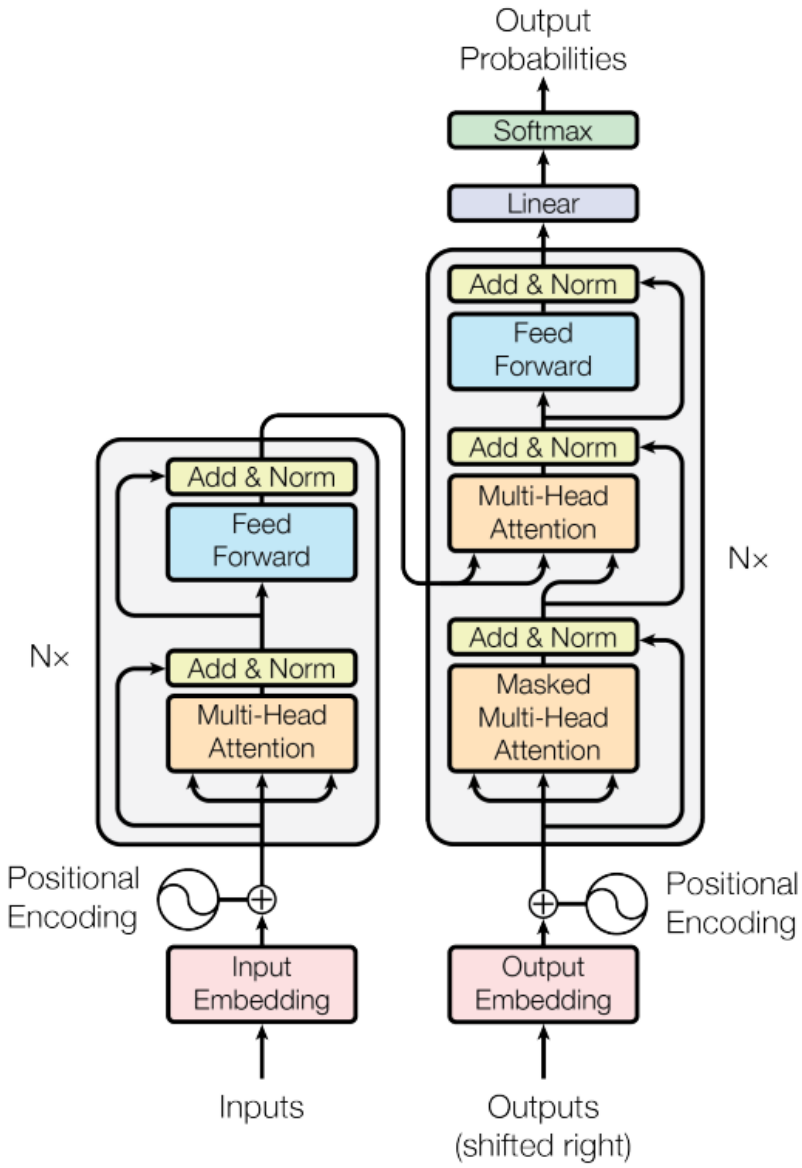

Figure 4.1: The full transformer architecture, taken from the original paper[113]. The left half is the "encoder" which generates a distributed representation of an input sequence, for example a sentence to be translated, and the right half is the "decoder" which uses that representation to generate the target output sequence, for example the translated sentence, in an auto-regressive manner.

needs some other mechanism to inject information about the absolute and relative position of the tokens in order to make use of the order of the sequence. This is the "Positional Encoding" unit shown in the diagram. We refer to the original paper[113] for the full details of multi-head attention and positional encoding, but remark here

that the number of "attention heads" is a hyperparameter indicating the number of different representation subspaces learned by each Multi-Head Attention block, on which the model jointly attends.

Apart from the attention mechanisms, the encoder and decoder also have a feed-forward layer which projects up to a higher dimensional space and then back down to the original. We refer to this factor by which the model projects upward as "forward expansion." As seen in the diagram, there is a residual connection[37] which goes around the attention and feedforward components, and a layer normalization module[5] after the residual connection. These combined elements constitute one encoder or decoder "layer," and the number of layers is a hyperparameter to be tuned.

4.1.1. Transformers on fMRI Data

In 2019, Devlin et al.[17] presented BERT, which stands for Bidirectional Encoder Representations from Transformers. Bidirectional meaning that each element of the sequence is able to attend both forward and backward in time. This is not always desirable, for example in language modeling, the next word to be generated is conditioned only on what has come before it. But in our case, we want the attention mechanism to learn as much as possible about the entire sequence. BERT removed the decoder half entirely during pretraining, using only a stack of encoder blocks which fed into an output layer. BERT was pretrained on a massive corpus of unlabelled text and the authors performed transfer learning by simply replacing the output layer to obtain state-of-the-art results on eleven natural language processing tasks. The simplicity of such transfer learning motivated us to adapt their approach. In recent years, this strategy of a stacked encoder transformer architecture has emerged as a superior alternative to recurrent methods for fMRI timeseries modeling. Bedel et al. (2023)[7] improved the state of the art for timeseries classification on multiple public fMRI datasets with a novel fused-window attention mechanism, but their work

did not explore pretraining or transfer learning. Nguyen et al. (2021)[70] achieved state of the art classification accuracy for a task-state decoding task on the Human Connectome Project 7-task dataset[112]. Their analysis includes the explicit benefits of the transformer's self-attention module when compared to previous recurrent architectures, as well as a demonstration of transfer learning when pretraining on held-out subsets of HCP 7-task. However, their pretraining task was supervised classification specific to HCP 7-task labelled data, and thus their pretrained models would be of little to no value toward transfer learning on different datasets or modalities[45]. Malkiel et al. (2022)[62] pretrain on a self-supervised fMRI reconstruction task by wrapping the transformer block in an encoder-decoder. They report that their pretraining was crucial for improved state of the art performance on a variety of fMRI tasks such as age and gender prediction, and schizophrenia recognition. We note that their downstream task uses the "CLS token" decoding method popularized by Devlin et al. (2019)[17], which we explain further below, and yet their pretraining task does *not* incorporate the CLS token. This inconsistency between training phases does not obtain the full value of the transfer learning paradigm.

Section 4.2

A Paired-Sequence Transformer for fMRI Tasks

Progressing from the previous section, we explored pretraining and transfer learning using a stacked encoder transformer model via novel self-supervised pretraining tasks which include the CLS token. All model inputs are in a standardized geographical brain space.

Our architecture is a modified stacked bidirectional-encoder design (Figure 4.2) with one or two separate output blocks, for each of two possible self-supervised pretraining tasks on which the model may be trained. We implemented our models

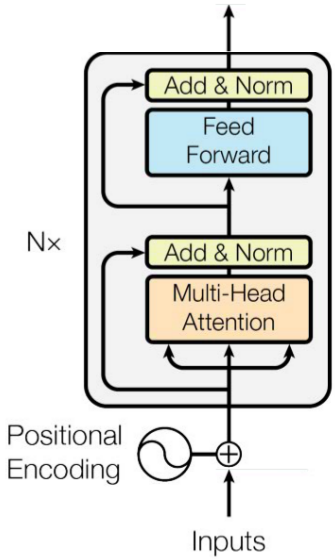

Figure 4.2: Our models consist of stacked transformer encoder layers without an embedding, and one or more output layers (not pictured). When we refer to the number of layers in a model, it is the N in this diagram being discussed.

from scratch with the pyTorch library. Our model does not include the standard embedding layer after positional encoding. We hypothesize that the composition of the fMRI scanner's measurement of BOLD signal with the mapping of that measurement to MNI space constitutes a meaningful embedding of the physical, biological neural response. The data are *already* in a shared, distributed, representative space. Hence, we dispense with the embedding layer in our design.

Each input to the model is constructed by extracting a contiguous sequence of fMRI images of a subject listening to music (**Seq1**), and then pairing it with another such sequence (**Seq2**). A **separator token** (**SEP**)[17] is inserted between the two sequences, and a **classification token** (**CLS**)[17, 62, 70] is inserted at the front. The presence of the SEP token in every input teaches the model to recognize the two separate sequences. CLS serves as a "pooling" token–since only the CLS token is fed to the output layer, backpropagation will force the model to learn to extract the necessary information from the rest of the input into the CLS token.

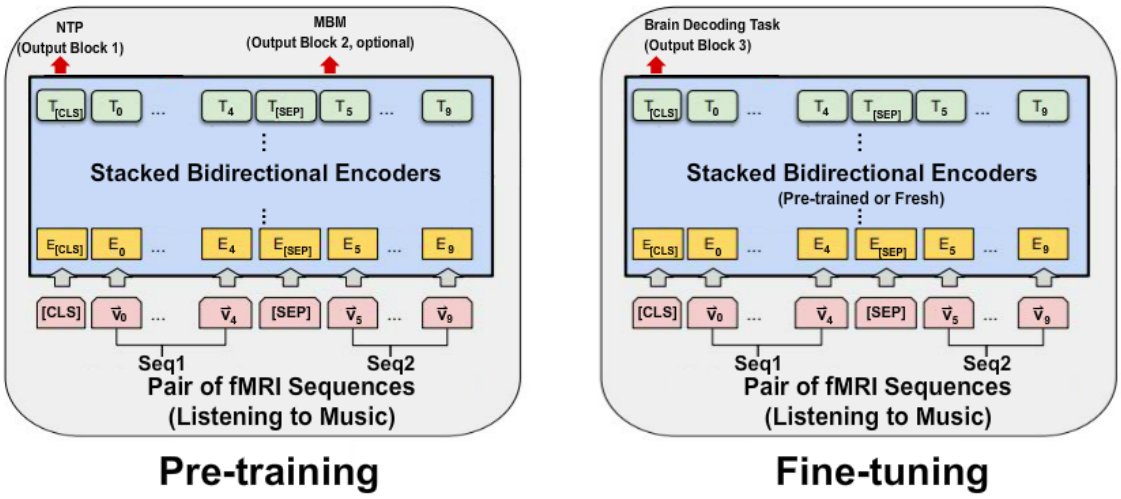

Figure 4.3: Pretraining and Finetuning phases. Output Blocks are not pictured but are detailed in corresponding sections. The model learns to extract information into the CLS token, which is fed to Output Block 1 during pretraining, and Output Block 3 during finetuning, for classification. The SEP token separates the two sequences. The masked token(s), if used, are fed to Output Block 2.

Before constructing the inputs, we reduced all fMRI images to only the left-side auditory cortex, resulting in 420 voxels, which we then flattened to 1-D. Thus each input x_i in the training set is a sequence of twelve 420-dimensional vectors, where Seq1 and Seq2 both have length N:

$$x_i = [CLS,\ \vec{v}_0,\ \ldots,\ \vec{v}_{N-1},\ SEP,\ \vec{v}_N,\ \ldots,\ \vec{v}_{2N-1}], \quad v_j \in \mathbb{R}^{420}. \tag{4.1}$$

In the domain of NLP, the above tokens are added as word indices to the vocabulary and the embedding layer learns their distributed representations[113]. Malkiel et al. (2022)[62] prepended a CLS token to sequences of fMRI images and pass that sequence through a learned embedding layer. But the original form of the CLS token must have the same dimension as the fMRI images in the sequence in order for the embedding layer to accept it. They do not report what this original form was. Logically, the tokens ought to be "far away" from the rest of the data in the distributed space. Thus

we simply reserved the first three of the 420 dimensions for our tokens. The CLS, SEP, and MSK tokens have a 1 in the zeroth, first, and second dimensions respectively, and are zero elsewhere. Each fMRI image has zero in those dimensions. Indeed, we had to remove the three voxels with lowest probability of inclusion from each image to make room for the tokens, and thus in practice only had 417 voxels in each image rather than 420. The success of our experiments, detailed in Chapter 5, **validates our novel implementation of these tokens without an embedding space**.

Section 4.3

Self-Supervised Pretraining

The classic supervised machine learning paradigm is based on learning in isolation, a single predictive model for a task using a single dataset. This approach requires a large number of training examples and performs best for well-defined and narrow tasks. **Transfer learning** refers to a set of methods that extend this approach by leveraging data from additional domains or tasks to train a model with better generalization properties[95]. Over the last few years, the field of Natural Language Processing (NLP) has witnessed the emergence of several transfer learning methods and architectures which significantly improved upon the state-of-the-art on a wide range of NLP tasks[17, 89]. This philosophy has begun to spread to deep learning on fMRI datasets[62, 70]. Following these works, we focus on *sequential* transfer learning, which consists of two phases: a **pretraining** phase in which general representations are learned on a source task followed by a **finetuning** phase during which the learned knowledge is applied to a target task.

Self-Supervised Learning (**SSL**), on the other hand, is a relatively new learning strategy which helps the model to learn universal knowledge based on a sort of pseudo supervision[95]. While supervised learning requires human labelled instances, the

labels in SSL are generated automatically based on data attributes and the definition of the task. This is what differentiates SSL from unsupervised learning, in which no labels whatsoever are used doing training. Moreover, as Kalyan et al. (2021)[45] describe, "the objective of unsupervised learning is to identify the hidden patterns while the objective of SSL is to learn meaningful representations."

The subsections below detail **our novel self-supervised pretraining tasks**.

4.3.1. Next Thought Prediction

Our first SSL task is Next Thought Prediction (**NTP**). The goal of NTP is binary classification, predicting whether or not Seq2 follows immediately after Seq1 in the original data. From the output of the final transformer block, the transformed CLS token is sent to **Output Block 1**. This block consists of a linear layer projecting down from 420 dimensions to 210, then a second linear layer projecting down from 210 to 2, and finally a softmax is applied to obtain probabilities for "No" (index 0) and "Yes" (index 1). The loss for NTP is calculated as the Cross-Entropy between the result of Output Block 1 and a one-hot encoding of the ground truth.

We remind the reader that Cross-Entropy is a measure of "difference" between two probability distributions over the same set. Here, the two distributions are over the output labels "No" and "Yes." Formally, the Cross-Entropy of the distribution q relative to a distribution p over a given set is defined as:

$$H(p, q) = -E_p[\log q], \tag{4.2}$$

where $E_p[\cdot]$ is the expected value operator with respect to p.

4.3.2. Masked Brain Modeling

Our second SSL task is Masked Brain Modeling (**MBM**). The goal of MBM is to reconstruct a masked element or elements of the input sequence. When an input arrives at the model, before positional encoding, 15% of the fMRI images in the input are chosen uniformly at random (without replacement) for masking. When an image is "chosen", there is an 80% chance to replace it with the **mask token** (MSK), a 10% chance to replace it with a random image sampled uniformly from the full dataset, and a 10% chance to leave it unchanged. These percentages are the same as in Devlin et al. (2019)[17]. The chosen indices are recorded, and the elements of the final transformer block's output at those indices are passed separately to **Output Block 2**. This block consists of a dense layer with ReLU activation, then a second dense layer with linear activation. The loss for MBM is calculated as the Mean Squared Error (**MSE**) between the output and the original chosen fMRI image. In the case of two chosen images, the total MBM loss is the average of the two individual MBM losses.

We remind the reader that MSE is a measure of the quality of a predictor, that is, a vector of n predictions. For each chosen fMRI image we have $n = 420$ output predictions ($\hat{Y}$) for the 420 true voxel values of that image (Y). Formally, the MSE of our MBM predictor is:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2. \tag{4.3}$$

Note the inherent data augmentation of the MBM task. There are ten fMRI images in each training sample, and the result of each possible masking configuration yields a distinct training sample. Thus MBM can effectively grow the size of the training set by an order of magnitude if the model is trained long enough. This gain is perhaps overlooked in domains such as natural language processing where billions of training samples are available. In fMRI studies, however, data poverty is a consistent concern due to the financial and time costs of scanning. We currently make no specific claims

about the effects of this augmentation in our experiments, but this potential benefit built into the task is noteworthy.

4.3.3. Multitask Learning

Training on more than one task simultaneously, known as **Multitask Learning**, has been shown to improve downstream performance in several domains[50] by benefiting from the underlying *relationships* between tasks, but to our knowledge this has not previously been done when training on fMRI data. In their thorough treatment of the brain's musical reward system, Salimpoor et al. (2015)[96] comment "music pleasure is thought to rely on generation of expectations, anticipation of their development and outcome, and violation or confirmation of predictions." In other words, the notion of "what comes next" is intimately connected to the explicit values of voxel activity. NTP and MBM embody these two concepts, so indeed our multitask pretraining scheme is aligned with the literature.

When training begins, the raw loss value of NTP for a single training sample is, on average, at least an order of magnitude greater than the loss value of MBM. Therefore the parameter updates will certainly be dominated by NTP, stifling any learning from MBM. Deriving a theoretically optimal way to combine the two losses would be a significant endeavor, so our total training loss for a single sample is merely the weighted sum of our two loss values:

$$E_{multi} = \alpha_1 \cdot E_{NTP} + \alpha_2 \cdot E_{MBM} \tag{4.4}$$

$$\alpha_1 + \alpha_2 = 1. \tag{4.5}$$

These two weights are simply hyperparameters to be tuned. We explore these and other hyperparameters in the Experiments and Results section below.

Section 4.4

Finetuning on a Brain Decoding Task

The second stage of sequential transfer learning is the **finetuning** phase, during which the learned knowledge from pretraining is leveraged toward a target task. The subsections below detail the target tasks, that is, our novel supervised brain decoding tasks.

4.4.1. Same-Timbre Task

The goal of the **Same-Timbre** task is binary classification on the Auditory Imagery Dataset detailed in Section 3.4.1, attempting to decode whether Seq1 and Seq2 correspond to the same timbre of music. Recall that each trial in this dataset is either Heard Clarinet (HC), Heard Trumpet (HT), Imagined Clarinet (IC), or Imagined Trumpet (IT). For this task, we consider HC and IC to be the same timbre, likewise for HT and IT. From the output of the final transformer block, the transformed CLS token is sent to Output Block 3. This block consists of a linear layer projecting down from 420 dimensions to 210, a second linear layer from 210 to 2 dimensions, then a softmax is applied to obtain probabilities for "No" (index 0) and "Yes" (index 1). The loss for Same-Timbre is calculated as the Cross-Entropy between the result of Output Block 3 and a one-hot encoding of the ground truth.

4.4.2. Same-Session Task

The goal of the **Same-Session** task is binary classification on the Enculturation Dataset, which was detailed in Section 3.5.2. Seq1 and Seq2 are either both Bach listening or both Shanxi listening, and the task is to predict whether the two sequences are drawn from the same scanning session. From the output of the final transformer block, the transformed CLS token is sent to Output Block 3. This block consists of a

linear layer projecting down from 420 dimensions to 210, a second linear layer from 210 to 2 dimensions, then a softmax is applied to obtain probabilities for "No" (index 0) and "Yes" (index 1). The loss for Same-Session is calculated as the Cross-Entropy between the result of Output Block 3 and a one-hot encoding of the ground truth.

4.4.3. Same-Genre Task

The goal of the **Same-Genre** task is binary classification on the Music Genre Dataset detailed in Section 3.4.2, attempting to decode whether Seq1 was and Seq2 were scanned while listening to the same genre of music. From the output of the final transformer block, the transformed CLS token is sent to Output Block 3. This block consists of a linear layer projecting down from 420 dimensions to 210, a second linear layer from 210 to 2 dimensions, then a softmax is applied to obtain probabilities for "No" (index 0) and "Yes" (index 1). The loss for Same-Genre is calculated as the Cross-Entropy between the result of Output Block 3 and a one-hot encoding of the ground truth.

Chapter 5

Experiments and Results

This chapter details the experiments we conducted with the BEAT framework presented in Chapter 4, as well as the results of those experiments. Section 5.1 presents the experiments and results on the Music Genre Dataset. Section 5.2 presents the experiments and results on the Auditory Imagery Dataset, as well as inference on the attention weights of the best performing models. Section 5.3 presents the experiments and results on the Enculturation Dataset, the collection of which was described in Chapter 3.

Section 5.1

Experiments on Music Genre Dataset

The Music Genre Dataset was detailed in Section 3.4.2. In this section we explain the construction of the training data, demonstrate our architecture's ability to learn the NTP and MBM pretraining tasks, including a potential benefit of training on them both simultaneously, and then demonstrate statistically significant benefits of transfer learning to the Same-Genre finetuning task.

As a helpful reference, we summarize the upcoming information about this dataset and its training data in Table 5.1.

Table 5.1: Summary of Music Genre Dataset. The Pretraining and Finetuning Samples fields are given as "{training samples} and {validation samples}".

Music Genre Dataset	
# of Participants	5
Training Regimen	12-fold cross-val, heldout runs
Region of Interest	Left STG
Pretraining Tasks	NTP+MBM Multitask, NTP-only
Pretraining Samples	10,000 and 800
Finetuning Task	Same-Genre
Finetuning Samples	10,000 and 800

5.1.1. Training Data

Recall that the stimulus schedule for this dataset was to present 15s clips of music from ten different genres with a TR of 1.5s, for a total of 10 consecutive fMRI images per clip with no break in between clips. Observe that the length of the inputs must be the same in both the pretraining and finetuning stages. Looking ahead to the Same-Genre task, the maximum length of Seq1 and Seq2 is therefore 10, because each clip of music is only 10 TRs long and crossing the boundary of another clip would create an input containing more than one genre. We note, though, that a small amount of boundary crossing, say 8 TRs of one clip and 2 TRs of the next clip with the 8 TRs deciding the genre of that sequence, could potentially make the learning more robust and lead to better generalization, similar to dropout. We leave that to future work.

Regardless, we instead chose a sequence length of 5 TRs to create more training data. We refer to this as a **5-seq** throughout the rest of this thesis. Another consideration is the *stride* when collecting 5-seqs. For example we could collect images 0 through 4, 2 through 6, and 5 through 9 to obtain three 5-seqs from a single clip. Or, we could avoid overlap entirely and collect only 0 through 4 and 5 through 9. Overlapping 5-seqs would have the advantage of more training data, at the cost of potentially overfitting the training data due to repeat exposure to certain images. We chose to

begin with no overlap at all (stride of 5), meaning two 5-seqs per clip, and met enough success that we have not, as yet, returned to consider different strides on this dataset.

With the above design, after collecting from all five participants we obtain 5400 5-seqs. Now the final consideration is the construction of the positive ("Yes" is the correct output) and negative ("No" is the correct output) training samples. Observe that it is impossible to have a positive sample for NTP if Seq1 and Seq2 are drawn from different participants. Thus each pairing, whether positive or negative, is drawn from a single participant, while the complete training set draws pairings from all participants, resulting in a sort of within/among-participants hybrid. Observe further that each 5-seq could be used as Seq1 for *both* a positive and negative sample when constructing the training set. This option has the advantage of doubling the amount of training data compared to choosing either a positive or negative partner for each Seq1. On the other hand, creating both could potentially contribute to overfitting from repeat exposure. For brevity going forward, we refer to this quality as "**PosNeg**." If PosNeg is True when when constructing a dataset, each 5-seq is used to construct both a positive and negative sample. If False, then each 5-seq is used to construct either a positive or negative sample with 50/50 chance.

The last consideration is the implementation of Masked Brain Modeling (**MBM**). Recall that the Masked Brain Modeling (**MBM**) masking process is performed when the input arrives at the model. Therefore creating a training set for multitask learning on both MBM and Next Thought Prediction (**NTP**) reduces to creating a training set for NTP-only and switching the masking process on or off. However, our goal was to select 15% of the fMRI images in an input for masking, but our inputs have two 5-seqs for a total of 10 images. Thus our implementation selects either one or two images with 50/50 change for an average of 15%. The masking process described in Section 4.3.2 is applied to the chosen images individually.

We are finally ready to construct the datasets. For pretraining, we performed 12-fold cross validation where each fold holds out as validation/test data of the twelve runs labeled as "Training" in the original dataset. We remind the reader that the original labels on the runs are not related to how we use them. For each fold, both the training and heldout data have stride of 5 and True PosNeg. The 5-seqs are paired according to the hybrid within/among scheme described above. The same 12 folds are used in all pretraining experiments below. For finetuning, we constructed 12 new folds with the same method as the pretraining folds.

5.1.2. Pretraining

One of the most important questions to ask in the context of multitask learning is whether the model would have been better off with only one task. In particular, *how much* is the performance on NTP impeded by having to learn MBM at the same time? To explore this, we performed our hyperparameter grid search for training on the multitask regimen as well as NTP alone. Recall that Nguyen et al. (2021) [70] was the only relevant transformer brain decoding work when we began these experiments. Thus we performed the initial hyperparameter search within small neighborhoods of their reported configuration (2 layers, 8 attention heads). This initial search met with enough success that we did not explore further. (Table 5.2 shows the best performing (i.e. achieved the highest validation accuracy on NTP at some point during training) configurations. We let the NTP task guide our search because its binary accuracy is simply more interpretable than any metric involving the MBM task. Nevertheless, the multitask models' performance on MBM is included in our analysis below.

All training during grid search held out run 0 from the dataset as a validation split. We applied a dropout rate of 0.1 in all transformer blocks. Models were trained for ten epochs via backpropagation with the Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.999$, and weight_decay $= 0.0001$.

Table 5.2: Best performing configuration for the two training regimens. Parameters from top to bottom are the alpha weights for loss calculation, learning rate, number of attention heads, and factor of forward expansion in the encoder blocks.

	Multi	NTP
α_1, α_2	0.1, 0.9	N/A
LR	10^{-4}	10^{-5}
Atn Hds	2	2
F Exp	4	4

In general, fewer attention heads with more layers outperformed the reverse. It is reassuring to obtain the same value for attention heads and forward expansion on both regimens. The best performing learning rate for NTP-only is an order of magnitude smaller than for multitask, but this is not surprising. NTP's contribution to the loss is scaled by $\alpha_1 = 0.1$, and in the most basic gradient descent, scaling the loss function by a constant is functionally the same as scaling the learning rate by that constant instead. The Adam optimizer is a bit more complex, but the general idea holds.

5.1.3. Cross Validation

After identifying the best performing hyperparameters for both cases, we performed 12-fold cross validation for both multitask and NTP-only, where each of the 12 folds held out one of twelve runs from the dataset. It was unclear during hyperparameter search whether a 3 or 4 layer model was superior, so we considered both here. The same Adam specifications as hyperparameter search were used. The exact details of pretraining dataset construction can be found in the Materials section below, but we note here that each fold has 10,000 training samples and 800 validation samples.

We applied a dropout rate of 0.1 in all transformer layers. Models were trained for ten epochs via backpropagation with the Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.999$, and weight_decay $= 0.0001$. Each model was trained with a different RNG seed for reproducibility. Results are presented in Table 5.3. For each fold, we saved the model's

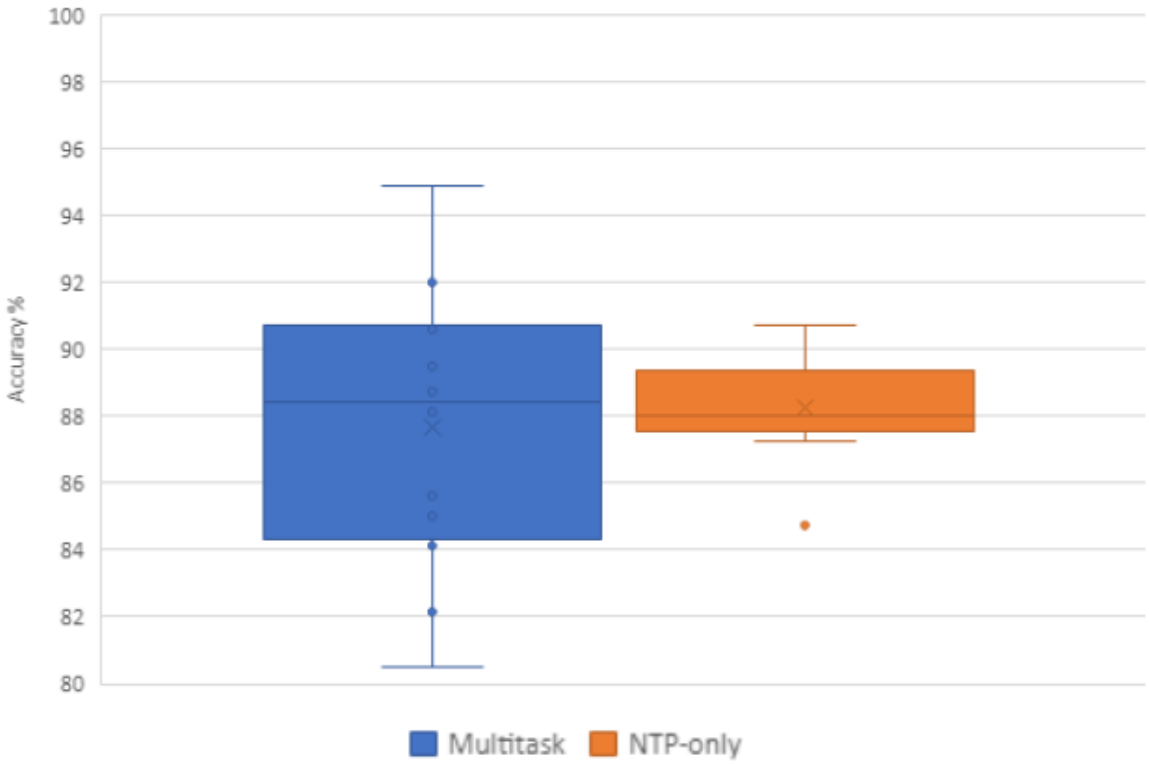

Figure 5.1: Box and whisker plot of the average Best Validation Accuracies obtained on the NTP task with two different pretraining strategies on the Music Genre Dataset: NTP-only, and both NTP and MBM simultaneously. Baseline chance on NTP is 50%.

state after the epoch with the highest NTP accuracy on the validation split ("Best Val. Acc." in the table). The "Best Epoch" column contains the epoch in which this accuracy was achieved. The MBM loss calculated on the validation split after the Best Epoch is also given to consider the relationship between the two tasks. Consider as a baseline that the MBM training loss on the first training sample seen by a model is around 0.4. The averages of each column are given in the last row of the table. The statistics of those averages are depicted in Figure 5.1.

Models with 3 layers outperformed on average on both accuracies of interest as well as MBM Validation Loss, so we proceeded to finetuning with the saved 3-layer models. The exact details of finetuning dataset construction can be found in the

Table 5.3: Results of 12-fold cross validation for Multitask (NTP and MBM) and NTP-only pretraining regimens, on 3 and 4 layers. Best Val Acc is the highest accuracy obtained during training on the NTP task on the validation split. Baseline chance on NTP is 50%. The epoch in which that accuracy was obtained is given in the Best Epoch column, from 0 to 9 inclusive. MBM Loss is the loss obtained on the MBM task on the validation split in the Best Epoch. The average across all twelve folds is given at the bottom of each column with ± standard deviation.

		Multitask (NTP and MBM)			**NTP Only**	
Heldout Run	**N Layers**	**Best Val Acc**	**Best Epoch**	**MBM Val Loss**	**Best Val Acc**	**Best Epoch**
0	4	93.5	8	0.00103	88.375	6
	3	88.125	8	0.00048	88.25	9
1	4	87.375	9	0.00088	87.375	8
	3	90.6	6	0.00051	88.375	9
2	4	88.625	4	0.00070	87.875	9
	3	88.75	9	0.00037	89.375	8
3	4	86.875	7	0.00043	87.375	8
	3	89.5	9	0.00118	87.75	7
4	4	80.0	3	0.00107	83.0	8
	3	80.5	8	0.00045	90.75	9
5	4	88.375	9	0.00080	87.0	9
	3	90.75	9	0.00040	87.75	9
6	4	79.375	8	0.00079	83.875	6
	3	84.125	8	0.00051	87.75	9
7	4	79.875	3	0.00259	85.375	9
	3	85.625	8	0.00071	89.25	9
8	4	81.75	6	0.00098	90.0	9
	3	94.875	8	0.00083	90.125	8
9	4	82.25	9	0.00102	85.0	8
	3	85.0	8	0.00076	84.75	4
10	4	80.375	5	0.00079	87.0	9
	3	92.0	9	0.00077	87.25	9
11	4	72.278	1	0.00070	88.734	9
	3	82.152	9	0.00032	87.468	9
Average	4	83.388±5.713	6	$(98 \pm 54) * 10^{-5}$	86.749±2.058	8.17
	3	87.613±4.255	8.25	$(61 \pm 25) * 10^{-5}$	88.237±1.557	8.25

Materials section below, but as in the pretraining phase, each fold has 10,000 training samples and 800 validation samples.

5.1.4. Finetuning

We loaded the twelve 3-layer models saved after their Best Epoch during the Multitask and NTP-only regimens and trained them for ten epochs on the Same-Genre task described above. The training data for each model holds out the same run as was held out during pretraining as a validation split. Preliminary testing showed that freezing the pretrained weights and updating only the new output block was not a successful training strategy for this work. Therefore all parameters were updated during finetuning. To examine the benefit of transfer learning, we also trained twelve randomly initialized (**RI**) models. The RI models are identical to the other models used in finetuning but do not load any pretrained weights.

The Adam optimizer parameters were the same as during pretraining. We trained all 36 models for 10 epochs with a Learning Rate of 10^{-4} and then again with 10^{-5}- the two learning rates used during pretraining. Table 5.4 reports the same values as Table 5.3 for the models with LR of 10^{-5}. The statistics of the average Best Validation Accuracies are depicted in Figure 5.2. These results outperformed the 10^{-4} results across the board so those are not reported.

5.1.5. Discussion

The first point of interest is that the pretraining phase was successful at all. fMRI data is a challenging domain and paired-sequence transformers have not previously been used on fMRI data, nor has multitask learning, in addition to our pretraining tasks being novel. Nevertheless, our implementation is conclusively capable of learning these tasks as both strategies significantly outperform the baseline chance of 50% (for each, a one-sample t-test against hypothetical mean of 50%, $p < .001$). The average best performance between the two regimens is not significantly different for the 3-layer models (paired t-test, p=.6661), which alleviates concerns about MBM impeding the ability to learn NTP. Moreover, it does not impede the *speed* at which

Table 5.4: Results of 12-fold cross validation for three finetuning regimens on the Same-Genre task: Multitask-pretrained models, NTP-only-pretrained models, and randomly initialized (RI) models, all with 3 transformer layers. Baseline chance on this task is 50%. Best Val. Acc. is the highest accuracy obtained during training on validation split. The epoch in which that accuracy was obtained is given in the Best Epoch column, from 0 to 9 inclusive. The average across all twelve folds is given at the bottom of each column with ± standard deviation.

	Multitask (NTP and MBM)		**NTP Only**		**RI**	
Heldout Run	Best Val. Acc.	**Best Epoch**	Best Val. Acc.	**Best Epoch**	Best Val. Acc.	**Best Epoch**
0	82.625	7	94.75	9	84.625	8
1	86.625	9	91.375	9	88.75	5
2	88.375	9	94.0	6	89.625	6
3	93.0	4	92.125	8	89.5	9
4	72.75	9	93.25	8	82.5	6
5	89.5	9	92.0	5	86.5	9
6	82.0	9	91.0	9	82.75	9
7	98.25	9	90.875	5	82.5	9
8	94.25	9	95.375	9	83.625	6
9	78.125	9	92.375	9	83.125	5
10	82.625	7	91.875	8	87.875	8
11	81.392	2	97.089	8	83.291	9
Average	85.793 ± 7.28	7.67	93.007 ± 1.93	7.75	85.389 ± 2.87	7.42

the multitask models achieve their best performance- about 8 epochs in both cases. The multitask models are more volatile, with standard deviation of 4.3 compared to 1.6 for NTP-only. NTP-only achieves its highest validation accuracy at 90.75%, but multitask runs achieve 92%, 93.5%, and 94.875%, which is our first evidence of a synergistic benefit from self-supervised multitask training on fMRI data.

Our novel supervised brain decoding task, Same-Genre, was also successful on both pretrained models and RI models. The models pretrained on NTP-only significantly outperformed the RI models, which is **our first significant evidence of the ability to perform transfer learning with our model from one of our novel self-supervised pretraining tasks to a supervised brain decoding task.** The

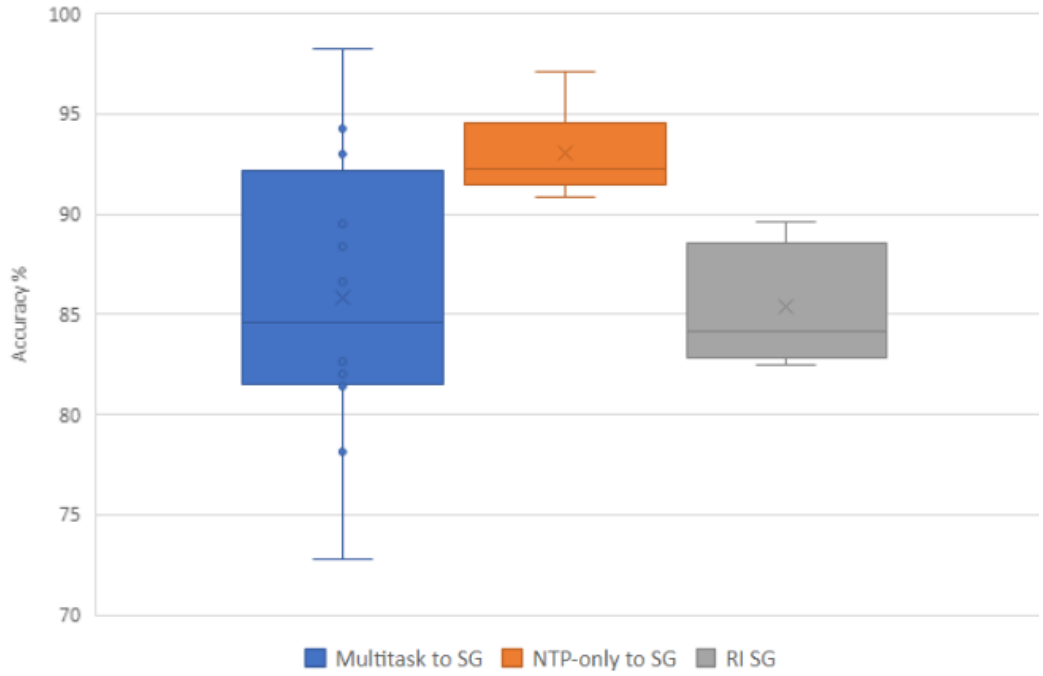

Figure 5.2: Box and whisker plot of the average Best Validation Accuracies obtained when learning the Same-Genre task with 12-fold cross-validation with three different initializations: pretrained on both NTP and MBM, pretrained only on NTP, and randomly initialized (RI). Baseline chance on this task is 50%.

models pretrained on Multitask almost exactly matched the baseline RI models on average, but we note a similar volatility to the pretraining phase. The average of the Multitask models is dragged down by folds 4 (72.75%) and 9 (78.125%). On the other hand, fold 7 achieves a staggering 98.25% validation accuracy, as well as 93% and 94.25%, all of which exceed the RI models' best fold of 91.625%. NTP-only reached a maximum of 97.089%, which is also short of the Multitask maximum. Note that the success in pre-training and fine-tuning also validates our novel implementation of the CLS and SEP tokens.

The relationship between pretraining performance and finetuning performance is unclear. For example, the second highest finetuning accuracy for Multitask was on folds 8, which was the highest performance of the 3-layer models during pretraining, indicating the positive relationship between the two phases that we would expect.

Table 5.5: Summary of Auditory Imagery Dataset. The Pretraining and Finetuning Samples fields are given as "{training samples} and {validation samples}".

Auditory Imagery Dataset	
# of Participants	17
Training Regimen	8-fold cross-val, heldout participants
Regions of Interest	Left and Right STG
Pretraining Task	NTP
Pretraining Samples	26,640 and 3,552
Finetuning Task	Same-Timbre
Finetuning Samples	2,520 and 336

On the other hand, fold 7 had the best performing Multitask finetuning accuracy, or rather the best finetuning accuracy of any regimen, while the pretraining accuracy and MBM loss were both below average on that fold. More work is required to properly identify a relationship between the two phases.

Due to time constraints we leave inference on the trained models of the Music Genre experiments to future work.

Section 5.2

Experiments on Auditory Imagery Dataset

The Auditory Imagery Dataset was detailed in Section 3.4.1. In this section explain the construction of the training data for NTP and Same-Timbre (ST, Sec.4.4.1). We then demonstrate our architecture's ability to learn the NTP pretraining task on this dataset as well as statistically significant benefits of transfer learning from NTP to ST in both Left and Right STG when generalizing to heldout subjects. Finally we briefly discuss our unsuccessful attempts at decoding explicit timbre labels of single-sequence inputs from this dataset. This shortcoming will reinforce our asking the question of decoding timbre in a different way, namely with ST.

We summarize the upcoming information about this dataset and its training data in Table 5.5.

5.2.1. Training Data

Recall that the source dataset consists of two each of Heard Clarinet, Heard Trumpet, Imagined Clarinet, and Imagined Trumpet. Holding out any single run as a validation set would not express the model's ability to learn all four conditions. If we hold out a strict subset from each of the four conditions, then some runs will contribute to both training and validation splits, raising data contamination concerns. Instead, we randomly shuffled the list of 17 participants and performed 8-fold cross validation, where fold n holds out all runs of subjects $2n$ and $2n+1$ in the shuffled list, $0 \leq n \leq 7$. This one shuffled list is the reference for constructing the 8 folds of all experiments on this dataset.

The pre-training data for NTP is created by selecting every possible 5-seq (with stride of 2) from all participants as Seq1, then for each of those, Seq2 is taken from the same participant as Seq1. As in the Music Genre experiments, our training data for NTP on this dataset is True PosNeg. Each fold during pretraining has 26,640 training samples and 3,552 validation samples. The training and validation splits are both half positive and half negative.

The finetuning data for Same-Timbre (ST) must match the architecture used in pretraining, so once again the inputs are pairs of 5-seqs. For this task, the 5-seqs are the five contiguous images beginning from the Target Note onset from each cycle (Figure 3.11). As in NTP, Seq2 was chosen within-participant. In all pairings, Seq1 and Seq2 are either both Heard or both Imagined. We leave the Heard-Imagined cross pairs to future work. As in pretraining, we construct both a positive and negative sample for each Seq1. However, observe the increased potential for data augmentation with this task. In NTP, each Seq1 has only one possible Seq2 to create a positive sample, but on this task the number of possible positive samples for each Seq1 is the total number of trials with that same timbre, which is 41. That is, we can augment

our ST datasets by a factor between 1 and 41 simply by the nature of having paired inputs. An ablation study on the effects of this augmentation is left to future work. In the Same-Timbre experiments below each of the 8 folds has 2,520 training samples and 336 validation samples.

The above datasets were constructed for both Left and Right STG separately.

5.2.2. Pretraining

We did not consider multitask pretraining on this dataset, instead focusing on the NTP-only regimen after seeing its superior transfer learning results on the Music Genre Dataset. We performed a basic hyperparameter grid search over the Learning Rate, the number of Transformer Layers, the number of Attention Heads within each layer, and the factor of Forward Expansion within each layer. The best performing configuration on the held out data was, in that order, [10^{-5}, 3, 2, 4]. Observe that this is identical to the optimal configuration found for NTP-only on the Music Genre dataset. This consistency contributes to the proof of concept of BEAT. We did not repeat hyperparameter search for any experiments below after observing this consistency and continued to use this configuration throughout.

We trained a model on the NTP task for each of the 8 folds described above. We applied a dropout rate of 0.1 in all transformer layers. Models were trained for ten epochs via backpropagation with the Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.999$, and weight_decay $= 0.0001$. Each model was trained with a different RNG seed for reproducibility.

Results for Left and Right STG are presented in Table 5.6. For each training session, we saved the model's state after the epoch with the highest NTP accuracy on the validation split ("Best Val. Acc." in the table). The "Best Epoch" column contains the epoch in which this accuracy was achieved, from 0 to 9. The averages of each column are given in the last row of the table. The statistics of those averages are

Table 5.6: Results of 8-fold cross-validation of pretraining on the NTP task with the Auditory Imagery Dataset. Baseline chance on this task is 50%. Sixteen of the seventeen participants were partitioned uniformly at random in groups of 2 to be held out as validation data for each of the 8 folds. Results for Left STG and Right STG are reported. Best Val Acc is the highest accuracy obtained during training on the validation split. The epoch in which that accuracy was obtained is given in the Best Epoch column, from 0 to 9 inclusive. The average of the best validation accuracies across the eight folds is given at the bottom of the corresponding columns with ± standard deviation.

	Left STG		**Right STG**	
Fold	Best Val. Acc.	**Best Epoch**	Best Val. Acc.	**Best Epoch**
0	87.6	5	91.4	6
1	84.8	7	87.5	6
2	84.7	4	86.1	9
3	88.3	6	86.7	4
4	90.9	9	92.3	8
5	83.7	4	89.6	8
6	87.9	9	87.2	7
7	88.2	8	85.3	9
Average	87.0 ± 2.4	6.5	88.3 ± 2.5	7.1

depicted in Figure 5.3.

All models on both hemispheres significantly outperformed the baseline chance of 50% when generalizing to heldout subjects (one sample t-test for both hemispheres against hypothetical mean of 50%, $p < .001$). However, there is no significant difference between the ability of Left and Right STG to generalize to heldout subjects (paired t-test between the two sets of accuracies, p=.2643).

5.2.3. Finetuning

For fine-tuning, the datasets were constructed for 8-fold cross-validation on the Same-Timbre (ST) task as described above, where each fold holds out the same subject as in pretraining to prevent data contamination. For each hemisphere (Left and Right), we finetuned by loading the saved pretrained weights from the eight Best Epochs of

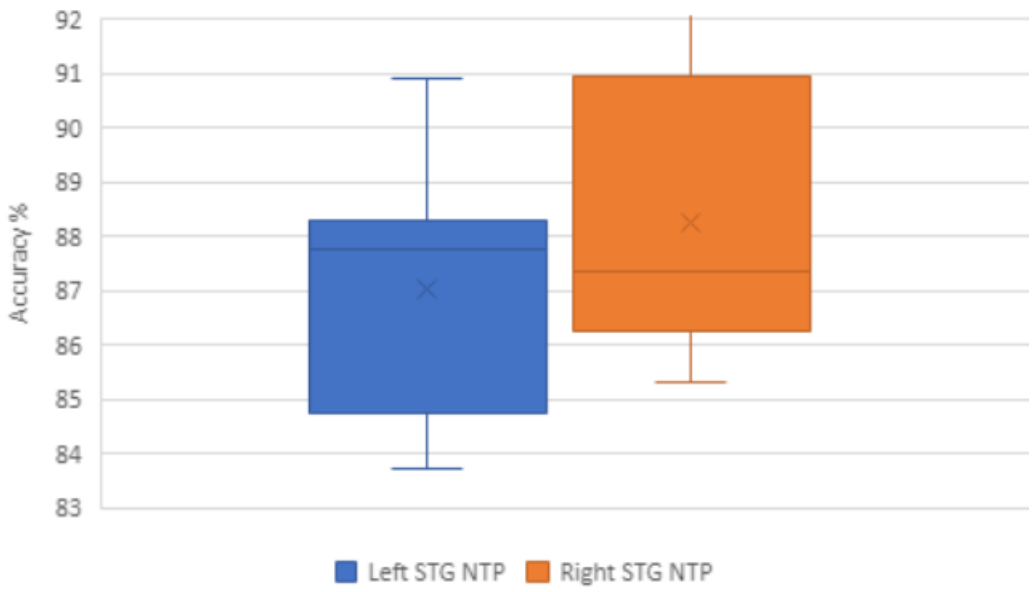

Figure 5.3: Box and whisker plot of the average Best Validation Accuracies for 8-fold cross-validation pretraining on NTP in Left STG and Right STG, using the Auditory Imagery Dataset. Baseline chance on this task is 50%.

pretraining on that hemisphere. The pretrained output layer was replaced with a single untrained Linear layer which projects from 420 dimensions down to 2, then a Softmax layer to obtain output probabilities for the "No" and "Yes" labels for ST. Preliminary testing showed that freezing the pre-trained weights and updating only the output layer was not a viable training regimen. Therefore all parameters were updated during fine-tuning. To examine the benefit of transfer learning, we also trained eight randomly initialized (RI) models with the same random-initialization parameters as in pretraining. The RI models are identical to the other models used in fine-tuning but do not load any pre-trained weights. We also trained eight null models on each hemisphere's data, which are identical to RI models but with the labels assigned uniformly at random to the training data. In the interest of brevity, we simply report here that the average best performance of the null models was roughly 51% in both hemispheres rather than reporting their individual results.

Models were trained for ten epochs via backpropagation with the Adam optimizer

Table 5.7: Results of finetuning with 8-fold cross-validation on ST in Left and Right STG after loading the Best Epoch weights from each fold in that hemisphere, as well as eight randomly initialized (RI) models which serve as a baseline to examine the effects of transfer learning in the two hemispheres. Baseline chance on this task is 50%. For each fold we report the highest accuracy obtained during training on that fold's heldout subjects, and the epoch in which it occurred from 0 to 9 inclusive. Note that this is not the epoch of the loaded pretrained weights, which can be found in Table 5.6. The average accuracy across all 8 folds is given at the bottom of the corresponding columns with ± standard deviation.

	Same-Timbre in Left STG				**Same-Timbre in Right STG**			
	Transfer from NTP		**RI**		**Transfer from NTP**		**RI**	
Fold	**Best Val. Acc.**	**Best Epoch**	**Best Val. Acc.**	**Best Epoch**	**Best Val. Acc.**	**Best Epoch**	**Best Val. Acc.**	**Best Epoch**
0	76.8	2	63.4	8	78.3	1	70.5	8
1	68.2	9	61.0	6	70.2	7	67.9	3
2	75.0	6	75.3	9	75.0	1	66.4	7
3	75.9	5	64.3	9	69.0	8	61.0	4
4	71.4	1	68.2	9	72.9	1	71.4	9
5	67.3	5	69.0	7	73.2	3	68.8	9
6	81.5	9	73.5	9	75.6	6	65.5	7
7	72.3	7	71.1	8	64.9	5	66.7	8
Avg.	73.5±4.7	5.5	68.2±5.0	8.1	72.4±4.2	6.9	67.3±3.2	6.9

with $\beta_1 = 0.9, \beta_2 = 0.999$, and weight_decay $= 0.0001$. Each model was trained with a different RNG seed for reproducibility. Table 5.7 contains our familiar metrics of Best Val. Acc. and Best Epoch. The statistics of the Best Val. Acc. averages are depicted in Figure 5.4.

All models significantly outperformed the baseline chance of 50% (one-sample t-tests, p¡.0001 in all cases). The Best Val. Acc. values obtained via transfer learning in Left STG are significantly higher than those obtained on the RI models (paired t-test between two sets of values, p=.0306). Similar for Right STG (paired t-test between two sets of values, p=.0105). There is no significant difference between Left and Right STG for the transfer learning models' performance (paired t-test, p=.5256).

Despite a seemingly lower average, the mean Best Epoch for Left pretrained models

Figure 5.4: Box and whisker plot of the average Best Validation Accuracies obtained when learning the Same-Timbre (ST) task via 8-fold cross-validation with four conditions: transferring from Left NTP to Left ST, learning Left ST with RI, transferring from Right NTP to Right ST, and learning Right ST with RI. Baseline chance on this task is 50%.

is not significantly different from the mean of Left RI models (paired t-test, p=.0691), and similar for Right (paired t-test, p=.1337). There is also no significant difference between the mean Best Epoch of Right and Left models, for either pretrained or RI (paired t-test, p=.1114 and .1501 respectively). Thus these experiments did not yield significant evidence of reduced training requirements due to transfer learning.

5.2.4. Direct Decoding and MVPA

Our paired-sequence approach to brain decoding is difficult to pitch in an elevator, so to speak, being entirely unfamiliar in this domain. Motivated by a desire to complement the above results with a more familiar approach, we attempted to use the same architecture to decode a single 5-seq rather than a pair of sequences. That is, the inputs to the model were length 6 and consisted of a CLS token and a 5-seq. All 5-seqs were extracted from STG from the dataset as for the ST task above. The CLS

Table 5.8: Preliminary hyperparameter search for decoding timbre directly from a single 5-seq with the CLS token. Each configuration has the same hyperparameters and architecture as the Same-Timbre experiments except for the hyperparameters listed here. Each configuration was trained for 20 epochs and the best accuracy obtained on the held out runs obtained by each configuration is reported along with the corresponding epoch number. We did not continue with this approach after all configurations significantly under-performed compared to the paired-sequence approach.

Encoder Layers	Attention Heads	Forward Exp.	Best Val. Acc.	Best Epoch
3	6	4	53.1	12
2	6	4	52.4	4
1	3	3	55.6	1
2	3	3	51.3	2
1	2	2	53.5	3

token was passed to the same output layer as above, with the binary classification corresponding to "trumpet" or "clarinet" rather than "true" or "false." The extent of our experiments with this approach was a preliminary hyperparameter search which clearly indicated that this approach was not viable. The results are given in Table 5.8. Each configuration was trained for 20 epochs and the highest validation accuracy across all epochs is given for each configuration, along with the epoch in which it occurred. Hyperparameters other than those shown in the table are the same as above. As shown in the table, all configurations significantly under-performed compared to our paired-sequence approach, and thus we did not continue with this method.

Prior to this work, during the research conducted for [64], our lab attempted to decode the clarinet and trumpet labels from both heard and imagined data using MVPA methods with an SVM classifier. Twenty regions of interest were considered. The regions which outperformed chance are shown in Table 5.9. The highest accuracy obtained was only 52.7%, and even though correction for multiple comparisons had yet to be done, none of the p-values were significant. Superior Temporal Gyrus is notably absent.

Table 5.9: MVPA methods with SVM classifier attempting to decode the clarinet and timbre labels. Twenty regions of interest were considered for both heard and imagined, and regions which failed to outperform chance are omitted. The highest accuracy is only 52.7%. These p-values had not yet been corrected for multiple-comparisons and were already insignificant.

ROI	H or I	Accuracy	P-value
RH Transverse Temporal	Heard	0.5266	0.0687
RH Banks STS	Imagined	0.5137	0.2502
LH Transverse Temporal	Imagined	0.5011	0.4787
RH Insula	Imagined	0.5028	0.4444

5.2.5. Discussion

Our pre-training phase on NTP was successful on this dataset as well. The models significantly outperforming chance on heldout runs complements our pre-training results from the previous section. We contribute this result as further significant evidence that the Next Thought Prediction task is well-defined and our novel paired-sequence architecture is capable of learning it.

Our novel supervised task, Same-Timbre, was also successful on Left and Right pre-trained models and RI models, while the null models performed only at random chance. We contribute this result as significant evidence that the Same-Timbre task is well-defined and our model is suited to learn it. Further, the pre-trained models significantly outperformed the RI models. We contribute this result as significant evidence of the ability to perform sequential transfer learning with our framework. Additionally, the success in pre-training and fine-tuning in these experiments further validates our novel implementation of the CLS and SEP tokens.

Observe that our implementation of the Same-Timbre task implicitly includes the task of decoding the Clarinet and Trumpet labels from a single 5-seq. For example, if we want to predict whether a Heard 5-seq is Clarinet or Trumpet, we need only pair it with a 5-seq whose Clarinet/Trumpet label is known, and then ask the model if

they are the same timbre. Most importantly, Same-Timbre significantly outperformed chance while our MVPA approach and direct decoding approach did not. Indeed, "asking the question in a different way" was needed for success. Thus we contribute our success on the Same-Timbre task as the first successful decoding of instrumental timbre from auditory cortex in fMRI data.

We take a closer look at the behavior of the attention mechanism and our hand-crafted tokens on this dataset's experiments in Section **??** below.

Section 5.3

Enculturation Dataset

Our motivation for collecting this dataset was detailed in Section 3.2.1. In the first set of experiments in this section we successfully trained our architecture to differentiate activity in Nucleus Accumbens between the first and second scanning sessions when the participant was listening to the Shanxi music, thereby establishing evidence of enculturation. We called this the Same-Session (SS) task. The second set of experiments demonstrate that NTP can be learned in Nucleus Accumbens using either the Music Genre Dataset (Genre NTP) or the Enculturation dataset (Enc NTP). The third set of experiments demonstrate transfer learning from both Genre NTP and Enc NTP to the Same-Session task for Shanxi listening.

We summarize the upcoming information about this dataset and its training data in Table 5.10.

5.3.1. Training Data

As discussed in Section 3.5.2, these experiments used the union of all participants' personal bilateral Nucleus Accumbens ROIs, consisting of 417 voxels, for a total input dimension of 420 after adding the three token dimensions, although MSK is currently

Table 5.10: Summary of Enculturation Dataset. The Pretraining and Finetuning Samples fields are given as "{training samples} and {validation samples}".

Enculturation Dataset	
# of Participants	5
Training Regimen	Fixed set of heldout runs, 12 iterations
Region of Interest	Union Nucleus Accumbens
Pretraining Tasks	Enc NTP and Genre NTP
Pretraining Samples	24,640 and 3,520 (Enc NTP)
	21,560 and 1,960 (Genre NTP)
Finetuning Tasks	Shanxi Same-Session, Bach Same-Session
Finetuning Samples	4,200 and 600 (both Bach SS and Shanxi SS)

not used. Arriving at the same input dimension as the prior experiments motivated us to being these experiments with the same hyperparameter configuration and training specifications which had been successful on the previous experiments. We encountered enough success that we have not yet revisited these decisions.

The training data for Genre NTP was constructed by extracting all 5-seqs from the twelve "Training" runs, as labeled in the original Music Genre Dataset, with stride of 2. Both a positive and negative NTP sample were constructed for each 5-seq to the left of the SEP token. All pairs were within-participant. 5-seqs whose positive NTP sample would cross a run boundary were not used to the left of the SEP token for positive or negative samples. In the Music Genre experiments above, the "Test" runs, as labeled in the original dataset, were included in our datasets in a tedious way with little benefit, so we excluded them here for simplicity. We obtained a total of 23520 samples for this task. During training, one of the twelve "Training" runs will be heldout for validation, resulting in 21560 training samples and 1960 validation samples.

The training data for Enc NTP was constructed similarly: all 5-seqs were used to make a positive and negative sample except for those which crossed a run boundary, a stride of 2, and within-participant pairings. The key difference is in the heldout

data. As always, we want to hold out entire runs as validation. But recall that each run of the Enculturation Dataset consists of four blocks, and each block has three trials. The three trials within a block are either all Bach or all Shanxi. Each full scan consists of 16 Bach blocks and 16 Shanxi blocks, but the arrangement of these blocks within a scan is randomized for each participant. For instance, an entire run may consist only of Bach blocks. Holding out such a run as validation would therefore bias the validation accuracy toward Bach and thus corrupt the results. By random chance, all five participants have at least one run that is two Bach blocks and two Shanxi blocks. If we were to hold out a run with blocks arranged as, say, Bach-Shanxi-Bach-Shanxi for all five participants, then we would bias the validation split toward that pattern of alternating musical grammars. Therefore we not only sought runs with 2 blocks of each style, but also with different arrangements of the four blocks across the five participants. Further, we wanted the run numbers to be as uniformly distributed as possible, to avoid bias toward the beginning or end of the scan. Thus we arrived at the following selection of heldout runs: Participant 1, run 3, Bach-Shanxi-Bach-Shanxi; Participant 2, run 7, Bach-Shanxi-Shanxi-Bach; Participant 3, run 6, Shanxi-Shanxi-Bach-Bach; Participant 4, run 4, Shanxi-Bach-Bach-Shanxi; Participant 5, run 1, Shanxi-Bach-Shanxi-Bach. Recall that the stimulus schedule for each participant was identical in sessions 1 and 2, so these runs are safe to hold out in both sessions. This process resulted in 24640 training samples and 3520 validation samples.

The training data for Same-Session was constructed differently. Each trial consists of 30 TRs, with padding on both ends, so for each trial we begin extracting 5-seqs at TR 4. We wanted to use a stride of 5 to avoid overfitting the training data via repeat exposure, so for each trial we extract the 5-seqs beginning at TRs 4, 9, 14, 19, and 24. With this task, we can create multiple positive and negative samples for each 5-seq,

unlike NTP, but in order to avoid overfitting the training data we only created one of each. For each input, both sides of the SEP token are either both Shanxi, or both Bach, and two different datasets were created for each style so their changes between sessions can be investigated in isolation. To avoid temporal correlation, the two sides of the SEP token are never drawn from the same block. The heldout runs were chosen as in Enc NTP for the same reasons. In particular, the uniform run numbers, as the participant's comfort in the scanner can change drastically over the course of a long scan, and thus either frontloading or backloading the heldout runs would introduce bias when analyzing NAcc. This process resulted in 4200 training samples and 600 validation samples for both Bach and Shanxi.

5.3.2. Same-Session Experiments

Our goal for these experiments was to obtain evidence of a musical enculturation effect in Nucleus Accumbens by training a classifier to predict whether two 5-seqs were drawn from the same session. We introduced this Same-Session (SS) task in Section 4.4.2. The construction of the training data was detailed in the preceding subsection.

The most intuitive results one might expect when training Same-Session on only Bach listening (Bach SS) is for the models to perform at random chance, since the experiment is not designed to change one's familiarity with western musical grammars. On the other hand, we expected statistically significant accuracy on the heldout data for Same-Session on only Shanxi listening (Shanxi SS) due to the changes in NAcc consequent to enculturation, as discussed in Section 3.2.1. Twelve models were trained on Bach SS, and another twelve on Shanxi SS. The results are given in Table 5.11. The highest accuracy on the heldout data is given in the Best Val. Acc. column, and the corresponding epoch in the Best Epoch column. The averages are given in the bottom row of the table, and the statistics of the Best Val. Acc. averages are depicted

Table 5.11: Results of training on the Same-Session (SS) task as detailed in Section 4.4.2 with heldout runs as detailed in Section 5.3.1 above. Baseline chance on this task is 50%. Twelve models were trained on pairs corresponding to Bach listening, and another twelve for Shanxi listening, indexed by the first column here. The highest accuracy on the heldout runs is given in the Best Val. Acc. column, and the corresponding epoch in the Best Epoch column. Epochs range from 0 to 9 inclusive. Averages are given in the bottom row with ± standard deviation.

	Shanxi SS		**Bach SS**	
Iteration	Best Val. Acc.	**Best Epoch**	Best Val. Acc.	**Best Epoch**
0	60.0	9	58.2	8
1	65.8	9	56.5	8
2	64.2	8	61.5	8
3	66.3	7	56.0	7
4	65.3	8	61.5	8
5	60.0	8	63.0	9
6	64.3	9	58.5	7
7	55.3	9	61.2	9
8	61.8	9	58.8	8
9	61.8	9	58.3	9
10	71.2	9	58.8	9
11	63.7	9	60.0	9
Average	63.7 ± 4.0	8.6	60.0 ± 2.1	8.3

in Figure 5.5.

Both sets of models are able to distinguish between sessions above the baseline chance of 50% (two one-sample t-tests with hypothetical mean 50%, $p < .001$ in both cases). We expected the Bach trials to be mostly indistinguishable between sessions as our control condition. One likely possibility for this distinguishability is the participants feeling more comfortable in the scanner the second time and being more familiar with the experiment. Observe though that the Shanxi trials are significantly more distinguishable than the Bach trials (paired t-test, p=.0096). Therefore if indeed the Bach distinguishability is due to the aforementioned confounds, then those confounds would also contribute to the distinguishability of Shanxi, yet we still have

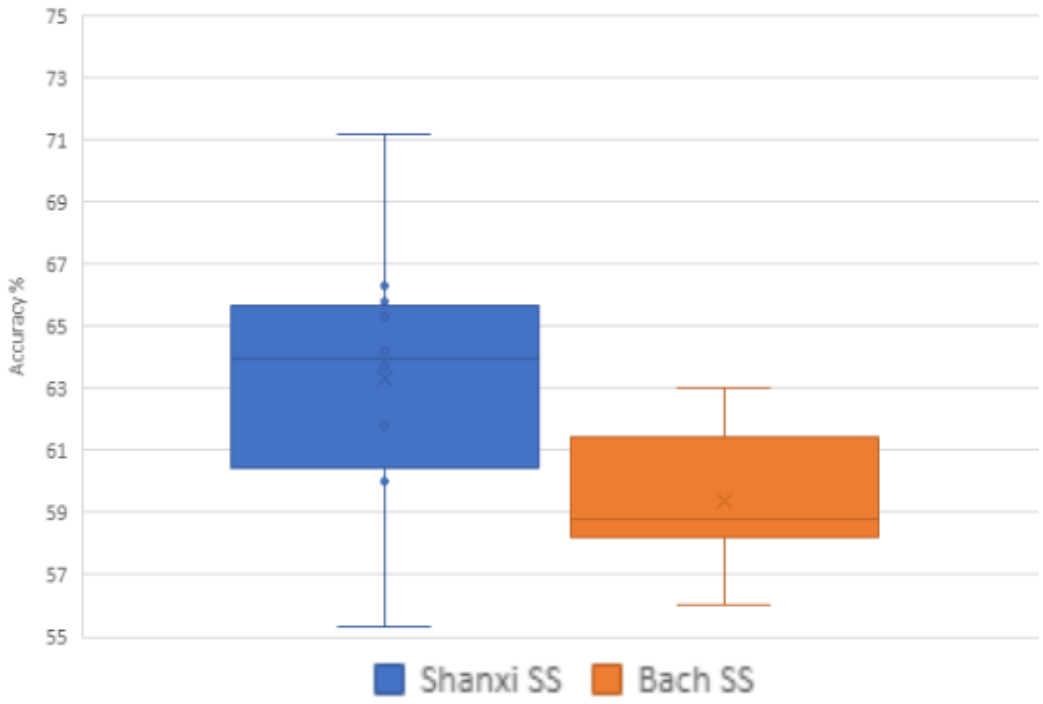

Figure 5.5: Box and whisker plot of the average Best Validation Accuracies obtained when learning the Same-Session task on just the Shanxi trials, as well as on just the Bach trials, with randomly initialized models. This served as a quick sanity check for the distinguishability of the two conditions, as we expected to see a greater distinguishability in the Shanxi trials after a week of at-home exposure. Baseline chance on this task is 50%.

a significant effect beyond those confounds. However, we argue here that neither the Bach nor Shanxi success is due to confounds such as familiarity and comfort.

Recall that these confounds were seen to modulate the amplitude of activity in the original EEG experiments discussed in Section 3.2.1. In all of our experiments (and fMRI research in general), the datasets are normalized to mean 0 and standard deviation 1. This should largely control for amplitude based effects due to familiarity or comfort, forcing the model to learn differences in patterns in the time series instead. We also directly tested for the presence of these sorts of confounds. If learnable effects due to familiarity and comfort in the scanner remain after normalization, we would expect these effects to be consistent across the entire scan, and thus both sets of

models (Bach SS and Shanxi SS) would be learning (at least partially) overlapping classifiers. In other words, the saved Bach models would recognize those confounds in the Shanxi data and perform with some amount of aptitude when evaluated on the Shanxi data, and vice versa. However, both cases yielded entirely degenerate behavior: when the Bach SS models were evaluated on Shanxi SS training data and vice versa, all models output "True" for all inputs, resulting in 50% accuracy. Indeed, there does not appear to be any overlap whatsoever between how the Bach and Shanxi models learn to distinguish the two sessions. This is compelling evidence that the models are not learning to recognize differences due to confounds such as familiarity or comfort. But then the models' aptitude for Bach SS remains unexplained. Prior to this work we hypothesized a potential anti-enculturation toward the western musical grammars, in which case it would be a *diminished* ability to make predictions on the western musical grammars that differentiates the two sessions. This could explain our results on Bach SS, but more work is required in this direction.

5.3.3. Pretraining

Our goal for pretraining was to determine whether our architecture can learn the NTP task in Nucleus Accumbens using both the Music Genre and Enculturation Datasets. Then, if so, save the best performing models to attempt transfer learning to Shanxi SS. For Genre NTP, we performed 12-fold cross-validation where each fold held out one of the twelve "Training" runs, as labeled in the original dataset. For Enculturation NTP, we trained twelve iterations with the heldout data constructed in Section 5.3.1 above, each with a unique RNG seed. The hyperparameters remain the same as in previous experiments.

The results are given in Table 5.12 and the familiar box and whiskers plot of the Best Val. Acc. averages is shown in Figure 5.6. Enc NTP and Genre NTP both outperformed the baseline random chance of 50% (two one-sample t-tests against

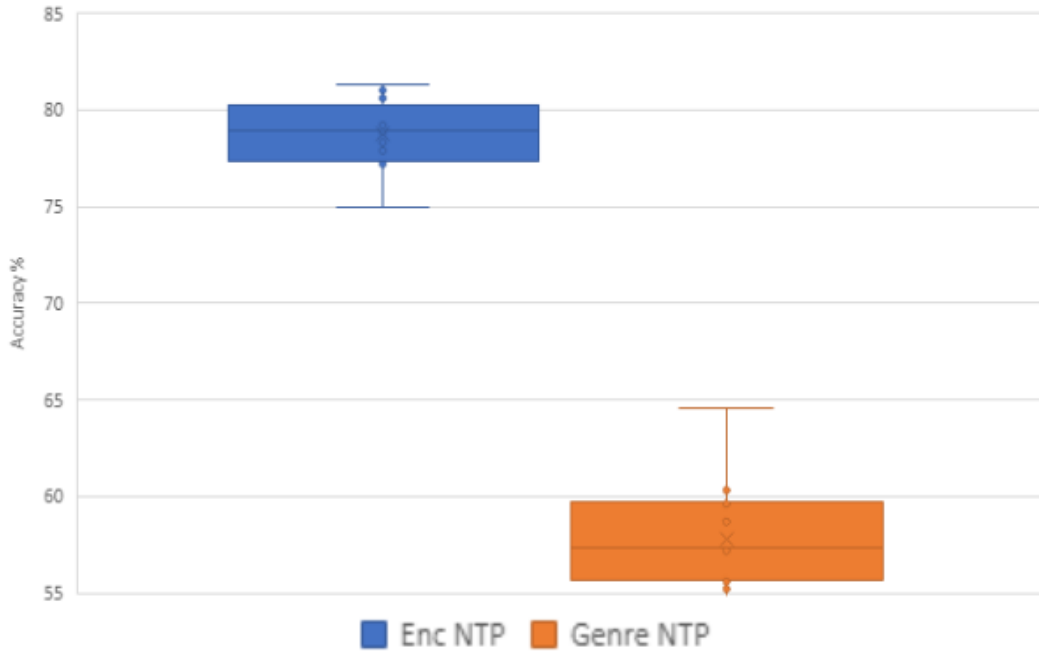

Figure 5.6: Box and whisker plot of the average Best Validation Accuracies obtained when pretraining on NTP in the Nucleus Accumbens ROI extracted from the Enculturation Dataset (Enc NTP) and the Music Genre Dataset (Genre NTP). Baseline chance on this task is 50%.

hypothetical mean of 50%, $p < .001$ in both cases). However, the Enc NTP models significantly outperformed the Genre NTP models (78.7 ± 1.8 and 57.8 ± 3.0 average Best Val. Acc. respectively, paired t-test with p¡.001), the latter of which also paled in comparison to our previous Genre NTP experiments in Superior Temporal Gyrus (88.2% average Best Val. Acc.). We hypothesize that there is a lower amount of variation in reward and prediction error consequent to the Music Genre stimuli as they all follow the familiar western musical grammars. This would make it more difficult to distinguish the temporal behavior of NAcc in the Music Genre participants. Nevertheless, these results contribute significant evidence that NTP can be learned outside of STG, in particular in NAcc, which reinforces our claim that NTP is a meaningful and well-defined self-supervised pretraining task for finetuning on fMRI brain decoding.

Table 5.12: Results of pretraining NTP on the Union Nucleus Accumbens ROI. Baseline chance on this task is 50%. We performed 12-fold cross-validation for the Music Genre dataset, with each fold holding out one of the twelve "Training" Runs, as labeled in the original dataset. Due to the nature of the stimuli design for the Enculturation runs, we cannot perform cross-validation with heldout runs, and instead trained twelve models with the heldout runs as explained earlier in this section. Each of these runs had a different RNG seed for reproducibility. Averages are given in the bottom row with ± standard deviation along with the epoch in which they occurred. Epochs range from 0 to 9 inclusive. Both experiments significantly outperform chance, although the Enculturation NTP models significantly outperform the Genre NTP models.

	Enc. NTP		**Genre NTP**	
Iteration	Best Val. Acc.	**Best Epoch**	Best Val. Acc.	**Best Epoch**
0	75.0	5	57.4	9
1	78.8	9	53.8	7
2	77.2	8	55.9	8
3	79.3	6	55.2	9
4	81.0	2	59.6	9
5	80.6	9	64.6	6
6	79.2	5	60.3	0
7	77.9	8	57.2	7
8	81.3	5	59.8	8
9	77.2	9	55.6	5
10	78.3	7	55.7	9
11	79.0	6	58.7	5
Average	78.7 ± 1.8	6.6	57.8 ± 3.0	6.8

5.3.4. Finetuning

Our goal for finetuning was to demonstrate transfer learning from both Enculturation NTP and Genre NTP to the Same-Session task. Due to time constraints, only Shanxi SS was examined. For each iteration of the NTP pretraining phases, the best performing models on the held out data (those reported in Table 5.12) were saved. For both NTP tasks, we loaded the twelve models, replaced the trained output layer with an untrained Linear layer which projects down from 420 dimensions to 2, and then a Softmax layer to obtain output probabilities for the two Same-Session classification

labels. The RNG seed for each iteration was identical to the above experiments as well, controlling out variation due to randomness when considering the benefit of transfer learning.

The results are shown in Table 5.13, and the corresponding averages are depicted in Figure 5.7. Inspection alone reveals that pretraining on Enculture NTP has a statistically significant benefit for accuracy on the heldout data (paired t-test between Best Val. Accuracies of Shanxi SS on Enc NTP and Shanxi SS on RI, p¡.001) as well as the speed of obtaining the maximum (paired t-test between Best Epochs of Shanxi SS on Enc NTP and Shanxi SS on RI, p¡.0001). These results are significant evidence of NTP being well defined in Nucleus Accumbens, and yet more evidence of its efficacy as a general self-supervised pretraining task for downstream brain decoding tasks on fMRI data. We note that iteration 11 when loading Enc. NTP weights is an outlier in its Best Epoch. However, after the first epoch this iteration obtained 68.7% accuracy on the heldout data. Thus this model did not need the extra time to outperform iteration 11 in the original Shanxi SS experiment, which had a Best Val. Acc. of 63.7%. Therefore we conclude that the shorter training requirement combined with the improved result was observed across all twelve iterations.

The models which were pretrained on Genre NTP are less dramatic, but nevertheless demonstrate significant improvements of Best Val. Acc. due to transfer learning compared to the RI models (paired t-test, p=.0416). To the best of our knowledge, this is the first significant evidence of the occurrence of transfer learning when the pretraining was performed on an fMRI dataset with wholly distinct participants and stimuli. Transfer learning across participants and stimuli would be necessary for a theoretical general pretrained model of the brain to be able to transfer its learning to arbitrary brain decoding tasks. Therefore we mark this result as the first major step toward that ultimate goal, and it is one of the core results of this thesis.

Table 5.13: Results of finetuning on the Shanxi Same-Session task. Baseline chance on this task is 50%. Finetuning was performed by loading models pretrained on either Enculturation NTP or Genre NTP. For each of the twelve iterations, the pretrained model was saved after the Best Epoch listed in Table 5.12. Epochs range from 0 to 9 inclusive. We include the RI Shanxi SS results from Table 5.11 for visual inspection of the transfer learning benefits. The Best Val. Acc. averages are given in the bottom row with ± standard deviation, as well as the Best Epoch averages.

	Shanxi SS on Enc. NTP		Shanxi SS on Genre NTP		Shanxi SS on RI	
Iteration	Best Val. Acc.	**Best Epoch**	Best Val. Acc.	**Best Epoch**	Best Val. Acc.	**Best Epoch**
0	66.0	2	62.2	4	60.0	9
1	69.5	1	63.0	10	65.8	9
2	67.0	6	63.3	4	64.2	8
3	64.7	3	65.2	5	66.3	7
4	71.7	3	70.7	4	65.3	8
5	70.2	2	64.2	9	60.0	8
6	71.8	1	69.0	10	64.3	9
7	70.2	3	65.7	8	55.3	9
8	70.3	2	68.7	5	61.8	9
9	69.7	1	63.7	9	61.8	9
10	70.0	3	69.2	9	71.2	9
11	72.3	10	66.3	10	63.7	9
Average	69.5 ± 2.4	3.1	65.9 ± 2.8	7.25	63.3 ± 4.0	8.6

5.3.5. Discussion

First, our models successfully learned the Same-Session task with performance on held-out runs significantly above chance, for both Bach and Shanxi listening. These results complement the EEG research conducted by Marion and Barbarot (Section 3.2.1) by demonstrating changes in Nucleus Accumbens as a result of musical enculturation. By evaluating the saved Bach SS models on the Shanxi SS data, and vice versa, we ruled out our worry about the models learning to differentiate the sessions by some global change due to familiarity or comfort. Moreover, these results reinforce two of our core contributions which we have been building up across this work: that our paired-input architecture is well-suited to fMRI based deep learning, and that our

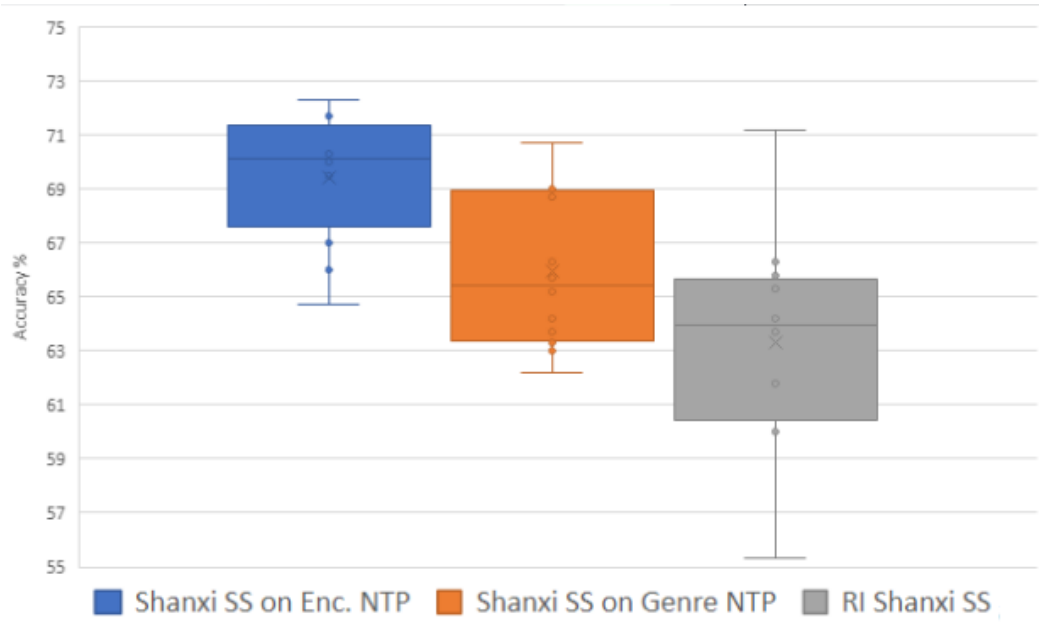

Figure 5.7: Box and whisker plot of the average Best Validation Accuracies obtained when performing transfer learning from Enc NTP and Genre NTP to Shanxi Same-Session, as well as the results of the RI Shanxi-SS models from Table 5.11 to examine the benefits of transfer learning. Baseline chance on this task is 50%.

novel implementations of the CLS and SEP tokens are effective.

Second, our models successfully learned the Next Thought Prediction task with the data drawn from Nucleus Accumbens, rather than STG as in the previous experiments. This was true for both the Enculturation dataset and the Music Genre Dataset. The Genre NTP models dramatically underperformed compared to both Enc NTP and the earlier Genre NTP on STG. We hypothesize that this is due to the Music Genre protocol consisting only of familiar western musical grammars, which results in less variation in the participants' internal prediction models.

Third, we performed transfer learning from both Enc NTP and Genre NTP to Shanxi SS, with statistically significant improvements in both cases. These results reinforce one of our core contributions, that NTP is a meaningful and effective self-supervised pretraining task for downstream brain decoding on fMRI data. They also provide another core contribution of this work, that it is possible to perform transfer

learning from one set of participants and stimuli to another, and more specifically that our framework and implementation are capable of it. To the best of our knowledge this is the first instance of such transfer learning on fMRI data.

Due to time constraints, we leave inference of the trained Enculturation models to future work.

Section 5.4

Inference

The above results have contributed significant evidence of BEAT's efficacy for self-supervised pretraining and transfer learning with fMRI data. We now take a closer look at *how* our models perform these tasks by observing the behavior of the attention weights. We will examine the top performing trained model from three of the above experiments: Left STG NTP, Left STG Same-Timbre after NTP pretraining, and Left STG Same-Timbre RI.

Recall that our models have three encoder layers, each with two attention heads, for a total of six. For a given input to the model, each attention head will calculate a 12x12 matrix of "attention weights," or "attention scores." Element i, j of this matrix represents how much attention element i of the sequence is paying to element j of the sequence. Each row in this matrix is the output of a softmax layer, and thus the rows are normalized and can be compared directly. The matrices are obtained by loading the best performing model from that experiment and setting the model to evaluation mode, in which no losses or gradients are calculated and no parameters are updated. We then input the full validation set that the model saw during training and we record the attention scores (the 12x12 matrix) obtained from each attention head. Because we are interested in how the model successfully performs each task, we only record the attention scores from inputs which the model classifies correctly. The collection of

12x12 matrices of attention scores is then averaged and presented below as heatmaps for the six heads in each of the three experiments. The coloring for each table is a linear gradient from fully red at its minimum value to fully green at its maximum.

5.4.1. Inference on Auditory Imagery NTP

The heatmaps for the six attention heads in this model are given in Figure 5.8. A critical and immediate sanity check is whether our implementations of the CLS and SEP tokens were able to convey to the model that the tokens are separate entities from the fMRI images. Indeed, the CLS and SEP columns are overwhelmingly red, with the curious exception of the SEP token attending on CLS in Layer 1 Head 2. It may be possible for the SEP token to serve as a secondary or auxiliary pooling site, which would explain why it attended on CLS. But then in Layer 2 SEP is soundly ignored so this explanation seems unlikely. The CLS token ignoring itself may seem alarming, after all it is intended to pool cumulative information across the three encoder layers. However, recall that there is a residual connection summing the input of the attention module to the output of the attention module, and thus the previous knowledge of the CLS token is not lost, but rather added to mostly new information in this case.

Observe the horizontal centralization of the higher attention scores. The model appears to have realized the most direct route to answering NTP is the relationship between $v4$ and $v5$, that is, the spot where the two sequences connect or not, resulting in lower attention scores on the fringes, where temporal relationships between the two sequences are harder to detect. This "shortcut" solution to NTP is undesirable, but note that this shortcut is not useful on any of our downstream brain decoding tasks. That is, the pairs in our brain decoding tasks will, by construction, never be connected at either end. Therefore the transfer learning benefits of NTP demonstrated in our results must be independent of this shortcut.

Next, observe the mostly red SEP columns directly in between the mostly green

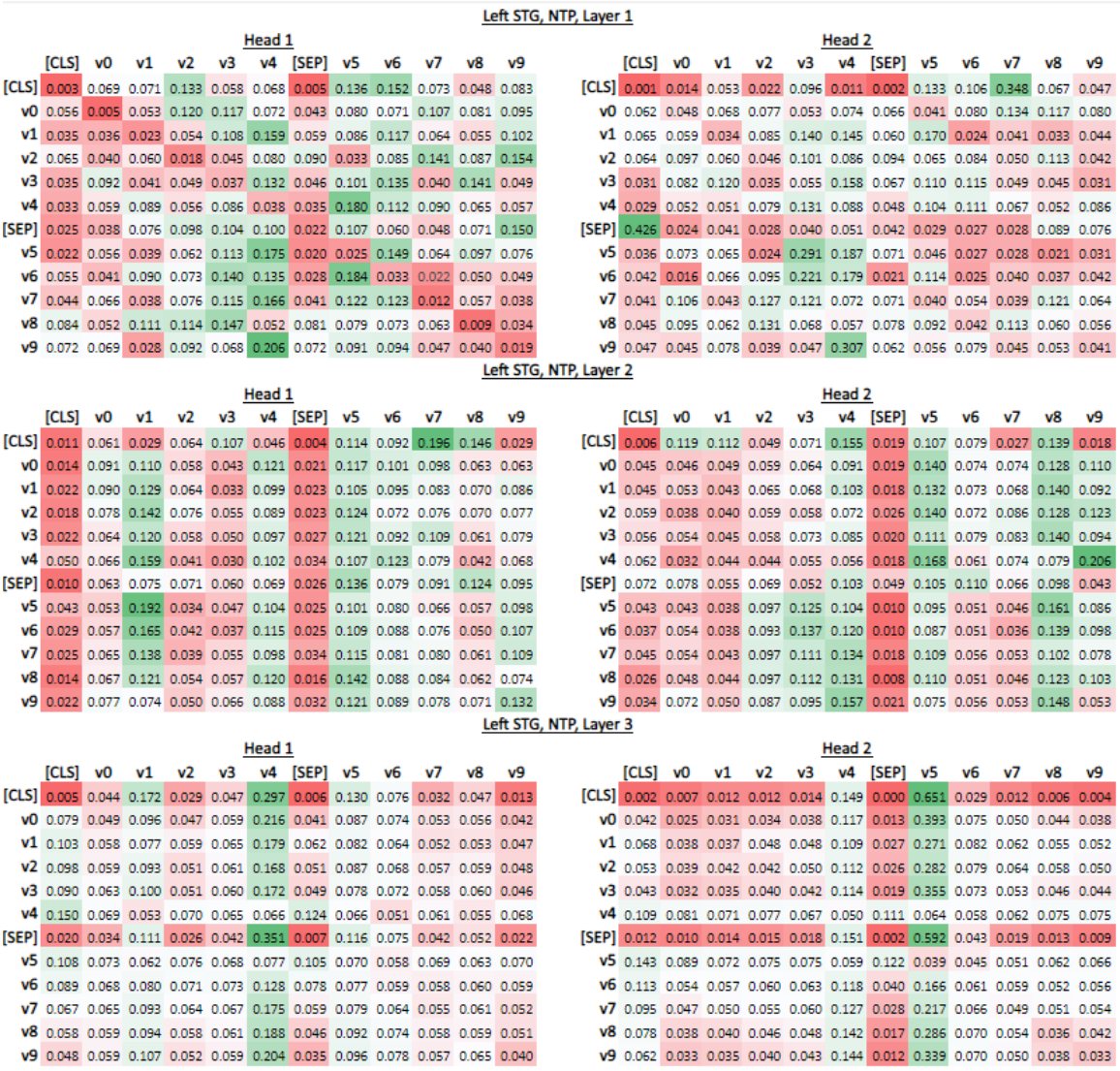

Figure 5.8: Heatmap of attention scores from red to green after averaging all correct validation inputs to the best performing saved model from NTP pretraining in Left STG of the Auditory Imagery Dataset.

columns of $v4$ and $v5$ in all six heads. This is only possible due to the Transformer architecture removing any elements of recurrence or relative position, instead including only information about absolute position. An LSTM or RNN, or even temporal-convolutional filters, would necessarily assume correlation between consecutive sequence elements, and would likely never obtain such a decisive decoupling of

important and unimportant neighbors.

Lastly for this model, recall that the output of the two attention heads in each layer are concatenated, and then observe the attention heads' division of labor between attending on $v4$ and $v5$ in Layer 3.

5.4.2. Inference on Same-Timbre RI

The heatmaps for the six attention heads are given in Figure 5.9. Compared to NTP, the best performing Same-Timbre RI model has a much smoother distribution of attention in Layers 2 and 3. However, this is a different task, and as mentioned above we wouldn't expect to see decisive attention on $v4$ and $v5$ this time. It is known, though, that the BOLD response peaks about 6 seconds after stimulus onset. The training data for Same-Timbre was constructed with the Target Note stimulus onset occurring on $v0$ and $v5$. The TR for this dataset was 2 seconds, and so it would be reasonable to expect the model to learn to focus on $v3/4$ and $v7/8$. This does not appear to have happened, as the model attends on all of the second sequence in general, with only two heads paying any attention to $v3$.

Note as well that all six heads have the fMRI images attending on the tokens. But perhaps we are being unfair, and Same-Timbre is just that much harder than NTP. Thus in order to properly evaluate these attentions, we must now consider the finetuned model.

5.4.3. Inference on Same-Timbre Transfer Learning

The heatmaps for the six attention heads are given in Figure 5.10.

Unlike the RI model, this model mostly avoids attending on the tokens, with the one exception being mostly green attention paid to the CLS token by the fMRI images in Layer 3 Head 1. While the persistence of this lesson from pretraining is a benefit, we also note the persistence of attending on $v4$ and $v5$, which is likely

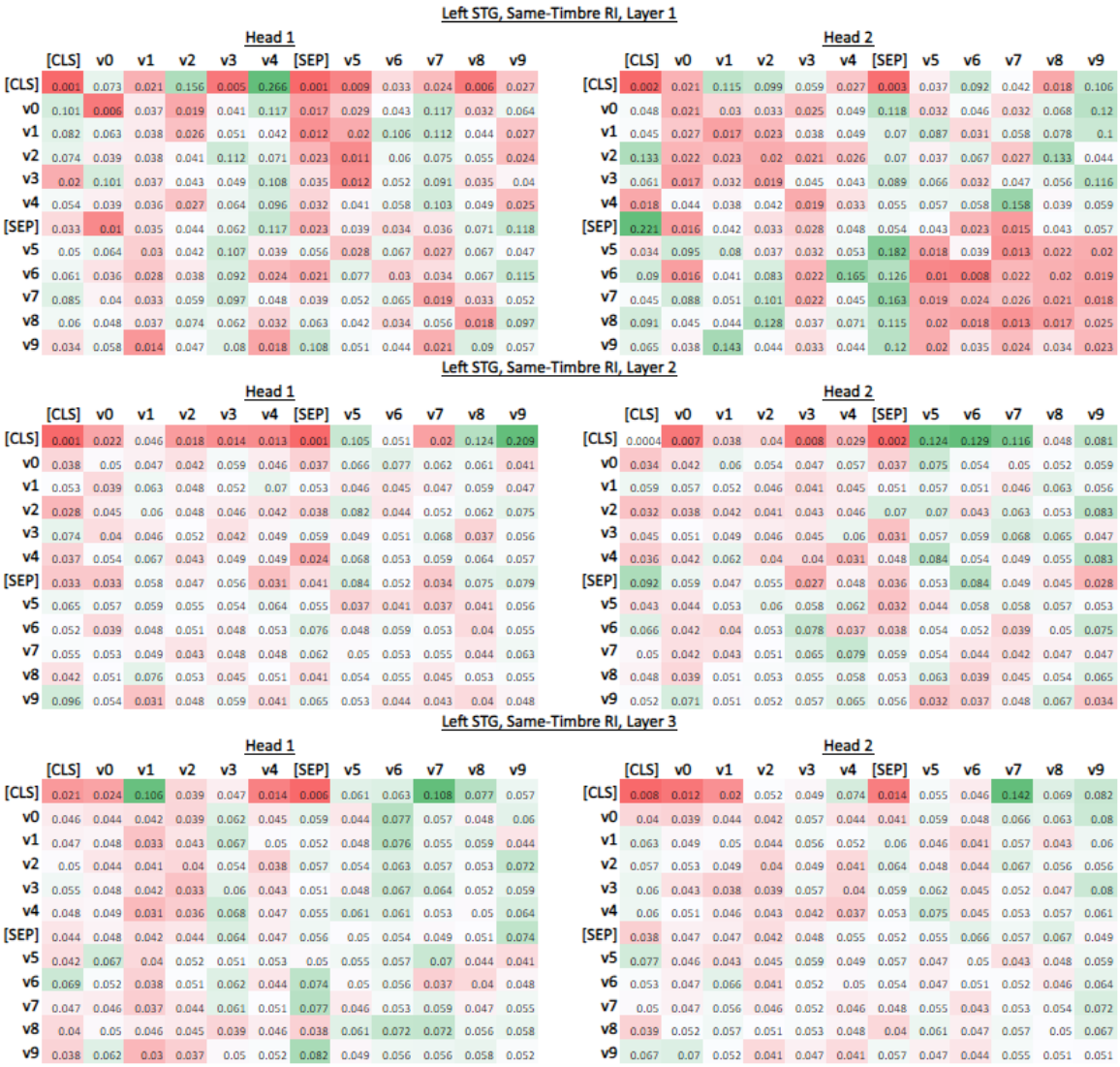

Figure 5.9: Heatmap of attention scores from red to green after averaging attention scores of all correct validation inputs to the best performing saved model from randomly initialized Same-Timbre in Left STG.

hindering performance by drawing attention away from $v3$ and $v7$. Similar inferences on our trained Music Genre and Enculturation models are left to future work due to time constraints. The yet unexplained attention on $v8$ in Layer 2 Head 2 persisted as well. Nevertheless the transfer learning models significantly outperformed the RI models on Same-Timbre, and thus we hypothesize that the most impactful benefit of

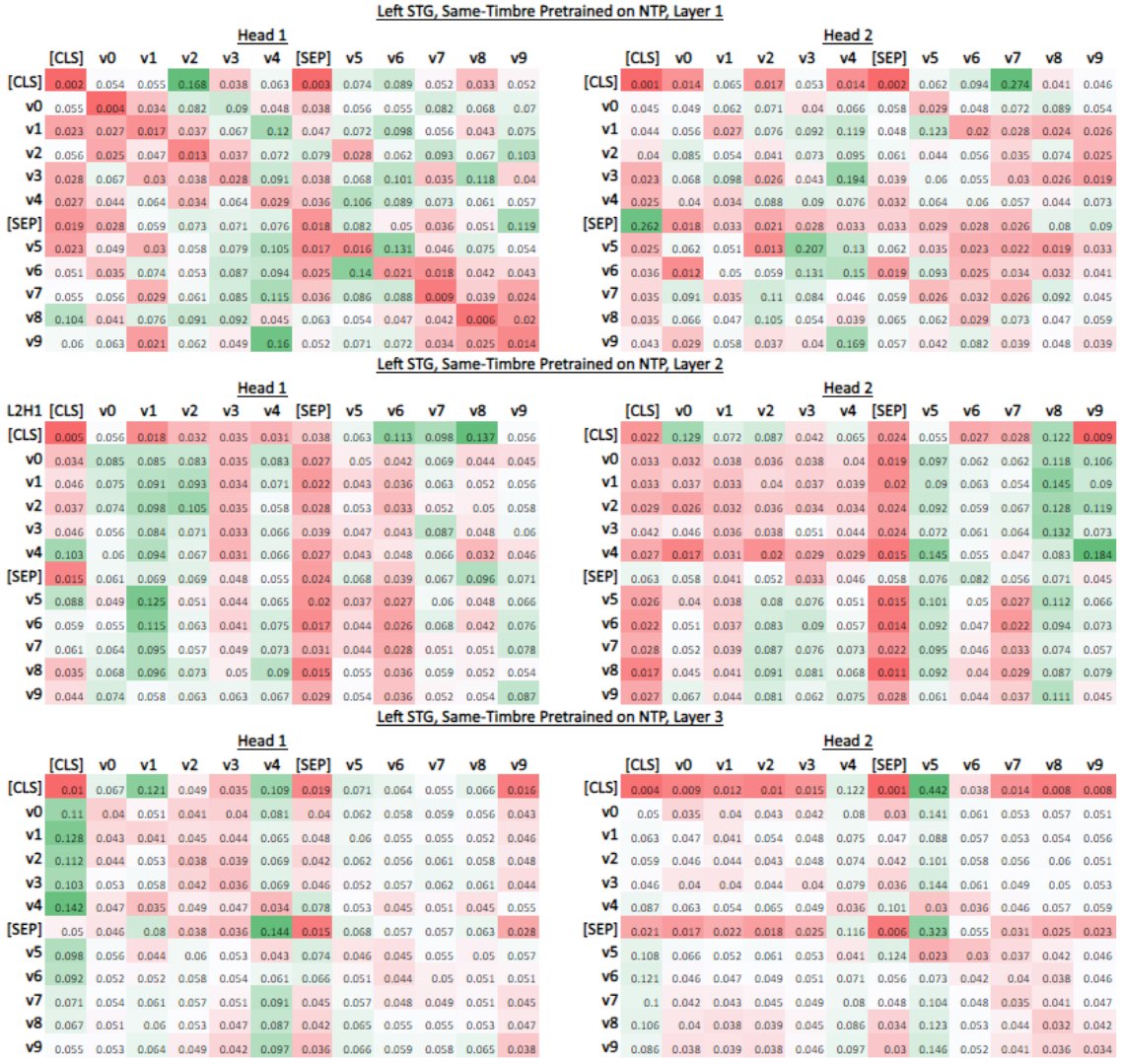

Figure 5.10: Heatmap of attention scores from red to green after averaging all correct validation inputs to the best performing saved model after transferring from NTP to Same-Timbre in Left STG.

NTP pretraining is learning not to attend on the tokens. These considerations can be developed further in future work with inference on the remaining trained models from the Auditory Imagery experiments, and further still on the models from the other experiments.

Chapter 6

Conclusions and Future Work

This chapter concludes this thesis. In Section 6.1 we summarize our core contributions. In Section 6.2 we discuss threads left dangling by this work and additional directions for progressing this area of research.

Section 6.1

Conclusions

Transfer learning is a powerful machine learning technique for improving downstream performance of deep learning models and reducing the demand for data. In our work immediately prior to the work in this thesis[79], we saw firsthand the necessity of learning the latent patterns in high dimensional fMRI data to improve performance of downstream classifiers. Thus we sought to develop a more formal transfer learning framework for the analysis of neural activity in fMRI data.

6.1.1. Contributions

In this work we presented BEAT, a novel sequential transfer learning framework for sequences of fMRI data. The contributions of BEAT are as follows:

- **Next Thought Prediction**: Self-supervised pretraining allows for leveraging

large amounts of unlabeled data, which is more readily available than labeled data, to learn useful representations. The pretraining phase captures general-purpose features that are likely to be useful for various downstream tasks. BEAT significantly outperformed chance when learning NTP on three different fMRI datasets (Genre, Auditory Imagery, and Enculturation), and in three different regions of the brain (Left and Right Superior Temporal Gyrus and Nucleus Accumbens). This is conclusive evidence that NTP is a well-defined self-supervised task and that BEAT has the capacity to learn it.

- **A transformer architecture without an embedding layer**: We hypothesized that an fMRI scan and subsequent preprocessing were functionally an embedding of physical cognitive processes into a representative distributed vector space, and thus the typical embedding layer was not needed. This required hand crafting the CLS, SEP, and MSK tokens in the distributed space. BEAT's success on all of our experiments supports this hypothesis and our implementation of the tokens.

- **Transfer learning**: BEAT obtained significantly improved performance on all three of our supervised brain decoding tasks–Same-Genre, Same-Timbre, and Same-Session–after pretraining on NTP. This contributes further conclusive evidence of the efficacy of NTP as a self-supervised pretraining task, but also contributes conclusive evidence that BEAT is well suited for facilitating transfer learning on sequences of fMRI data. The development of such a framework was the primary goal of this thesis. In the case of Same-Session, BEAT obtained significant benefits when pretrained on a different dataset entirely, thus achieving the first step toward a generalized pretrained model of the brain, which was the secondary goal of this thesis.

- **Brain decoding results**: Our attempts to decode instrumental timbre from STG with MVPA methods and with a direct decoding method were unsuccessful. However, BEAT's successful learning of the Same-Timbre task contributes significant evidence that instrument-specific features are represented in both Left and Right STG and can be distinguished. BEAT's successful learning of the Same-Session task contributes significant evidence of a musical enculturation effect in Nucleus Accumbens after a week of exposure to an unfamiliar musical grammar.

6.1.2. Limitations

Recall that each fold of Same-Timbre training had only 2,520 training samples and Same-Session had only 4,200 training samples. BEAT was only able to learn these tasks because we restricted the data to relatively small regions of interest. If larger areas of the brain are of interest, concerns with data poverty begin to emerge. On the other hand, both of the aforementioned tasks can benefit from the natural data augmentation of paired tasks. We discuss this more in the next Section. So indeed, while BEAT does lack the ability to output explicit labels for arbitrary single-sequence or single-image brain decoding tasks, the data augmentation is potentially quite powerful. More importantly, our work was singularly focused on the question of *distinguishability* of conditions in the brain, from which we draw conclusions about the representations of those conditions, e.g instrumental timbre in STG. Explicit label decoding is not always necessary for modern brain decoding goals, and this observation hearkens back to Haxby (2012) [34] in which the author credits part of the success of MVPA to the asking of questions in a different way.

Masked Brain Modeling faces limitations in theory as well as in implementation. For the theory, consider Tong et al. (2022) [109]. In the context of masking and reconstructing sequences of pixel-regions in video, they remark that the temporal

correlation of the video frames will lead to data leakage, allowing the models to learn "shortcut" features that will not generalize well. fMRI data has high temporal correlation, so this concern applies to MBM as well. Perhaps MBM learned temporally correlated shortcut features to reconstruct the masked images, rather than gaining a high-level understanding. This hypothesis offers an explanation for our multitask pretraining performing as well as NTP-only, and yet failing to transfer its knowledge to Same-Genre, while NTP-only transferred with statistical significance. In implementation, the limitation comes from the task being asked of the model. Mean Squared Error Regression to a 420 dimensional vector in a space known for its complex subtleties is unlikely to approach meaningful reproduction. As Lu et al. (2023) [61] note, the most powerful models working on this task have been able to reconstruct images that are semantically similar to the originals, but "the outcomes are always lacking in control over details such as location and size." We discuss alternatives to our MBM implementation in the Future Work section below.

A major limitation in the collection of our Enculturation Dataset was the human-element. Our scanning protocol was roughly 1.5 hours per scan, which is intimidating for any potential recruits. Further, scheduling fMRI scans exactly one week apart when our scanner is under constant heavy demand is difficult, and we were not entirely successful as two participants had their second scan after an extra day. Marion and Barbarot exposed their participants to two weeks of Shanxi music, but our scheduling and recruitment challenges would have only magnified if we tried to match their two weeks of exposure. With only half the desired enculturation period, the effects may not have developed enough, resulting in less distinguishable sessions. This likely contributed to BEAT's overall weaker performance on Same-Session compared to our other brain decoding tasks, not to mention the fact that only one of the five participants listened to the *minimum* requested amount of music on all days (30 minutes), while

none of them listened to the *recommended* amount every day (Table 3.1).

Section 6.2

Future Work

As our primary interest is the efficacy of transfer learning, there are a considerable number of ablation studies still to explore to optimize downstream performance. These include:

- Different constructions of the training data: stride, threshold for inclusion in ROI, among-participant pairing for downstream tasks, etc. In particular, examination of the benefits of data augmentation available in paired sequence tasks, for example, multiplying the dataset for the Same-Timbre task by a factor of up to 41.

- Additional regions of interest for experiments on Enculturation Dataset: ventromedial prefrontal cortex, dorsal striatum, right angular gyrus, etc. Refer to Section 3.2.1 for more.

- BEAT performed transfer learning from Genre NTP to Same-Session in NAcc. A natural follow-up, then, is pretraining on Enc NTP in STG and attempting to transfer the learning to Same-Genre.

- In the Same-Timbre experiments, all pairs were either both Heard or both Imagined. However, in our previous work[79], we were able to train our SVM classifier on Heard and then perform above chance when evaluating on Imagined– the "cross-decoding" result. It remains to consider cross-decoding experiments with BEAT.

- One particular strength of the transformer architecture is the ability to learn

long-term dependencies, and it remains to repeat our work with longer sequences. This requires specific scanning protocols, however. For example, with the Genre Dataset, the maximum length would be 10 TRs before samples would contain music of more than one genre. Indeed, it may be hard to find fMRI datasets with stimuli or tasks long enough to fully explore learning long sequences.

In Masked Language Modeling, the model output is a vector of probabilities over the vocabulary space. The index with the highest probability is therefore the model's prediction for the masked word. The model does not actually need to reconstruct the distributed (embedded) representation of the masked word. This stands in contrast to MBM, which, as discussed in the Limitations section above, suffers from the very difficult task of reconstructing a complex high dimensional vector. Vector Quantized Variational-Autoencoders (VQ-VAEs) [92, 77] present an exciting alternative to our implementation of MBM. In short, VQ-VAEs learn discrete representations of images, and have been deployed successfully on 256x256 images [92], much larger than the ROIs in this work. Imagine we have a trained VQ-VAE for an fMRI dataset, and the model is about to replace an image in the input sequence with the MSK token. We would then make a quick detour to obtain the chosen image's discrete representation, that is, some integer, and record it for later. The transformer blocks then proceed as usual. But the output layer can now perform a different task, similar to MLM, we can now output a probability distribution over the space of discrete representations (integers), with the previously recorded integer as the ground truth. This is a much simpler task than 420 dimensional regression. Further, it will mitigate or even eliminate our concerns about data leakage. Since the model is no longer trying to reconstruct the voxel values, the temporal correlations of the voxels in temporally-local images likely cannot provide the "shortcuts" discussed in the Limitations section above. Continuing with this area of thought, we could also implement an fMRI analogue of standard

Language Modeling, in which the task is to predict the next work. Attempting to generate the next fMRI image in voxel space faces the same challenges as MBM with regression, but VQ-VAE would provide a similar alternative wherein the model predicts the discrete representation of the next image instead.

Our extraction of STG from the Auditory Imagery and Music Genre Datasets via the HO Atlas and FSLeyes may fall short of current standards for Regions of Interest in fMRI data, and Future Work ought to consider more advanced methods such as the Glasser [28] or Schaefer [98] parcellations.

Additional work is required to identify the features BEAT learned in order to distinguish the two scanning sessions of Bach trials despite our experiment having been designed with that as our control.

An examination of the correlates of musical features with evoked BOLD signal could grant insight into the behavior of BEAT's attention heads, as perhaps the model learns to attend on musical features encoded in the data.

While the use of the standardized MNI space generally accounts for different spatial resolutions between source and target data, different temporal resolutions, that is, different TR times, must be addressed to progress toward a generalized pretrained model of the brain. Sample Rate Conversion is a ubiquitous and critical function of signal processing systems[129] and we are interested in the potential for transfer learning when these techniques have been applied to convert the temporal resolution of a target dataset to that of the pretraining data.

We would like to continue to show the ability of BEAT to learn the NTP task on various regions of the brain scanned with various protocols, and transfer that learning to downstream brain decoding tasks, particularly non-audio tasks, as so far we have only used audio-evoked fMRI data.

The most important direction of Future Work, however, is pretraining on larger and

larger datasets to improve transfer learning from one dataset to another. Pretraining is generally thought of as having an enormous set of unlabelled training data, while in our experiments we have thus far only used fairly typical single-study sized datasets for pretraining. The benefit obtained on the Same-Session task by pretraining on Genre NTP was relatively weak and only barely statistically significant, thus we would like to improve this effect with a larger pretraining dataset. With our long-term goal of a generalized pretrained model of the brain, we need to obtain more knowledge from more participants during pretraining. The Human Connectome Project[112] is the most compelling avenue towards this goal and is the most immediate line of research we propose to pursue, due to its overwhelming size and well-established benchmarks across the entire brain.

Bibliography

[1] Reginald B. Adams and Petr Janata, *A Comparison of Neural Circuits Underlying Auditory and Visual Object Categorization*, NeuroImage **16** (2002), no. 2, 361–377 (en).

[2] Kat Agres, Samer Abdallah, and Marcus Pearce, *Information-Theoretic Properties of Auditory Sequences Dynamically Influence Expectation and Memory*, Cognitive Science **42** (2018), no. 1, 43–76 (en).

[3] Arafat Angulo-Perkins, William Aubé, Isabelle Peretz, Fernando A. Barrios, Jorge L. Armony, and Luis Concha, *Music listening engages specific cortical regions within the temporal lobes: differences between musicians and non-musicians.*, Cortex; a journal devoted to the study of the nervous system and behavior **59** (2014), 126–137 (eng).

[4] John Ashburner, *Preparing fMRI Data for Statistical Analysis*, fMRI Techniques and Protocols (Massimo Filippi, ed.), vol. 41, Humana Press, Totowa, NJ, 2009, Series Title: Neuromethods, pp. 151–178.

[5] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton, *Layer Normalization*, (2016), Publisher: arXiv Version Number: 1.

[6] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, *Neural Machine Translation by Jointly Learning to Align and Translate*, (2014), Publisher: arXiv Version Number: 7.

[7] Hasan Atakan Bedel, Irmak Şıvgın, Onat Dalmaz, Salman Ul Hassan Dar, and Tolga Çukur, *BolT: Fused Window Transformers for fMRI Time Series Analysis*, February 2023.

[8] C Buchel, *The functional anatomy of attention to visual motion. A functional MRI study*, Brain **121** (1998), no. 7, 1281–1294.

[9] César Caballero-Gaudes and Richard C. Reynolds, *Methods for cleaning the BOLD fMRI signal*, NeuroImage **154** (2017), 128–149 (en).

[10] Vincent K.M. Cheung, Peter M.C. Harrison, Lars Meyer, Marcus T. Pearce, John-Dylan Haynes, and Stefan Koelsch, *Uncertainty and Surprise Jointly Predict Musical Pleasure and Amygdala, Hippocampus, and Auditory Cortex Activity*, Current Biology **29** (2019), no. 23, 4084–4092.e4 (en).

[11] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation*, (2014), Publisher: arXiv Version Number: 3.

[12] David D Cox and Robert L Savoy, *Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex*, NeuroImage **19** (2003), no. 2, 261–270 (en).

[13] R Cameron Craddock, G Andrew James, Paul E Holtzheimer, 3rd, Xiaoping P Hu, and Helen S Mayberg, *A whole brain fMRI atlas generated via spatially*

constrained spectral clustering, Hum Brain Mapp **33** (2011), no. 8, 1914–1928 (en), Place: United States.

[14] Jumana Dakka, Pouya Bashivan, Mina Gheiratmand, Irina Rish, Shantenu Jha, and Russell Greiner, *Learning Neural Markers of Schizophrenia Disorder Using Recurrent Neural Networks*, December 2017.

[15] Anders M. Dale and Randy L. Buckner, *Selective averaging of rapidly presented individual trials using fMRI*, Human Brain Mapping **5** (1997), no. 5, 329–340 (en).

[16] Steven M. Demorest, Steven J. Morrison, Denise Jungbluth, and Münir N. Beken, *Lost in Translation: An Enculturation Effect in Music Memory Performance*, Music Perception **25** (2008), no. 3, 213–223 (en).

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, May 2019.

[18] Giovanni M Di Liberto, Claire Pelofi, Roberta Bianco, Prachi Patel, Ashesh D Mehta, Jose L Herrero, Alain De Cheveigné, Shihab Shamma, and Nima Mesgarani, *Cortical encoding of melodic expectations in human temporal cortex*, eLife **9** (2020), e51784 (en).

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, June 2021.

[20] Nicha C. Dvornek, Pamela Ventola, Kevin A. Pelphrey, and James S. Duncan, *Identifying Autism from Resting-State fMRI Using Long Short-Term Memory*

Networks, Machine learning in medical imaging. MLMI (Workshop) **10541** (2017), 362–370 (eng).

[21] Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio, *Why Does Unsupervised Pre-training Help Deep Learning?*, Journal of Machine Learning Research **11** (2010), no. 19, 625–660.

[22] Oscar Esteban, Christopher J Markiewicz, Ross W Blair, Craig A Moodie, A Ilkay Isik, Asier Erramuzpe, James D Kent, Mathias Goncalves, Elizabeth DuPre, Madeleine Snyder, and others, *fMRIPrep: a robust preprocessing pipeline for functional MRI*, Nature methods **16** (2019), no. 1, 111–116, Publisher: Nature Publishing Group.

[23] Bruce Fischl, *FreeSurfer*, Neuroimage **62** (2012), no. 2, 774–781, Publisher: Elsevier.

[24] Elia Formisano, Federico De Martino, Milene Bonte, and Rainer Goebel, *"Who" Is Saying "What"? Brain-Based Decoding of Human Voice and Speech*, Science **322** (2008), no. 5903, 970–973 (en).

[25] K. J. Friston, A. P. Holmes, K. J. Worsley, J.-P. Poline, C. D. Frith, and R. S. J. Frackowiak, *Statistical parametric maps in functional imaging: A general linear approach*, Human Brain Mapping **2** (1994), no. 4, 189–210 (en).

[26] K.J. Friston, A.P. Holmes, J-B. Poline, P.J. Grasby, S.C.R. Williams, R.S.J. Frackowiak, and R. Turner, *Analysis of fMRI Time-Series Revisited*, NeuroImage **2** (1995), no. 1, 45–53 (en).

[27] Bruno L. Giordano, Stephen McAdams, Robert J. Zatorre, Nikolaus Kriegeskorte, and Pascal Belin, *Abstract Encoding of Auditory Objects in Cortical Activity Patterns*, Cerebral Cortex **23** (2013), no. 9, 2025–2037 (en).

[28] Matthew F. Glasser, Timothy S. Coalson, Emma C. Robinson, Carl D. Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F. Beckmann, Mark Jenkinson, Stephen M. Smith, and David C. Van Essen, *A multi-modal parcellation of human cerebral cortex*, Nature **536** (2016), no. 7615, 171–178 (en).

[29] R Goebel, B Sorger, J Kaiser, N Birbaumer, and N Weiskopf, *BOLD brain pong: Self regulation of local brain activity during synchronously scanned, interacting subjects*, 2004.

[30] Benjamin P. Gold, Ernest Mas-Herrero, Yashar Zeighami, Mitchel Benovoy, Alain Dagher, and Robert J. Zatorre, *Musical reward prediction errors engage the nucleus accumbens and motivate learning*, Proceedings of the National Academy of Sciences **116** (2019), no. 8, 3310–3315 (en).

[31] Andrea Halpern, *Differences in Auditory Imagery Self-Report Predict Neural and Behavioral Outcomes*, Psychomusicology: Music, Mind and Brain **25** (2015), 37–47.

[32] Michael Hanke, Florian Baumgartner, Pierre Ibe, Falko Kaule, Stefan Pollmann, Oliver Speck, Wolf Zinke, and Jörg Stadler, *A high-resolution 7-Tesla fMRI dataset from complex natural stimulation with an audio movie*, Scientific Data **1** (2014).

[33] Erin E. Hannon and Sandra E. Trehub, *Tuning in to musical rhythms: Infants learn more readily than adults*, Proceedings of the National Academy of Sciences **102** (2005), no. 35, 12639–12643 (en).

[34] James V. Haxby, *Multivariate pattern analysis of fMRI: the early beginnings*, NeuroImage **62** (2012), no. 2, 852–855 (eng).

[35] James V. Haxby, M. Ida Gobbini, Maura L. Furey, Alumit Ishai, Jennifer L. Schouten, and Pietro Pietrini, *Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex*, Science **293** (2001), no. 5539, 2425–2430 (en).

[36] Soichi Hayashi, Bradley Caron, Anibal S. Heinsfeld, Sophia Vinci-Booher, Brent C. McPherson, Daniel N. Bullock, Giulia Berto, J. Guiomar Niso, Sandra Hanekamp, Daniel Levitas, Lindsey Kitchell, Josiah Leong, Filipi N. Silva, Serge Koudoro, Hanna Willis, Jasleen Jolly, Derek Pisner, Taylor Zuidema, Jan Kurzwaski, Koulla Mikellidou, Aurore Bussalb, Christopher Rorden, Conner Victory, Dheeraj Bhatia, D. Baran Aydogan, Frank C. Yeh, Franco Delogu, Javier Guaje, Jelle Veraart, Jeremy Fischer, Joshua Faskowitz, Maximilien Chaumon, Ricardo Fabrega, David Hunt, Shawn McKee, Shaw T. Brown, Stephanie Heyman, Vittorio Iacovella, Amanda Mejia, Daniele Marinazzo, Cameron Craddock, Emanuele Olivetti, Jamie Hanson, Paolo Avesani, Eleftherios Garyfallidis, Daniel Stanzione, James P. Carson, Robert Henschel, David Y. Hancock, Craig A. Stewart, David Schnyer, Damian Eke, Russell A. Poldrack, Nathalie George, Holly Bridge, Ilaria Sani, Winrich Freiwald, Aina Puce, Nicholas Port, and Franco Pestilli, *brainlife.io: A decentralized and open source cloud platform to support neuroscience research*, 2023.

[37] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, *Deep Residual Learning for Image Recognition*, (2015), Publisher: arXiv Version Number: 1.

[38] Elizabeth M. C. Hillman, *Coupling mechanism and significance of the BOLD signal: a status report*, Annual Review of Neuroscience **37** (2014), 161–181 (eng).

[39] Sepp Hochreiter and Jürgen Schmidhuber, *Long Short-Term Memory*, Neural Computation **9** (1997), no. 8, 1735–1780 (en).

[40] Seyed Hani Hojjati, Ata Ebrahimzadeh, Ali Khazaee, Abbas Babajani-Feremi, and Alzheimer's Disease Neuroimaging Initiative, *Predicting conversion from MCI to AD using resting-state fMRI, graph theoretical approach and SVM*, Journal of Neuroscience Methods **282** (2017), 69–80 (eng).

[41] Tomoyasu Horikawa and Yukiyasu Kamitani, *Generic decoding of seen and imagined objects using hierarchical visual features*, Nature Communications **8** (2017), no. 1, 15037 (en).

[42] Zhenghui Hu and Pengcheng Shi, *Interregional Functional Connectivity via Pattern Synchrony*, 9th International Conference on Control, Automation, Robotics and Vision, 2006, ICARCV '06, January 2007, pp. 1 – 6.

[43] Heng Huang, Xintao Hu, Yu Zhao, Milad Makkie, Qinglin Dong, Shijie Zhao, Lei Guo, and Tianming Liu, *Modeling Task fMRI Data Via Deep Convolutional Autoencoder*, IEEE transactions on medical imaging **37** (2018), no. 7, 1551–1561 (eng).

[44] Petr Janata, *The Neural Architecture of Music-Evoked Autobiographical Memories*, Cerebral Cortex **19** (2009), no. 11, 2579–2594 (en).

[45] Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha, *AMMUS : A Survey of Transformer-based Pretrained Models in Natural Language Processing*, August 2021.

[46] Yukiyasu Kamitani and Frank Tong, *Decoding the visual and subjective contents of the human brain*, Nature Neuroscience **8** (2005), no. 5, 679–685 (en).

[47] C Krumhansl, *Cross-cultural music cognition: cognitive methodology applied to North Sami yoiks*, Cognition **76** (2000), no. 1, 13–58.

[48] M. Kubicki, R. W. McCarley, P. G. Nestor, T. Huh, R. Kikinis, M. E. Shenton, and C. G. Wible, *An fMRI study of semantic processing in men with schizophrenia*, NeuroImage **20** (2003), no. 4, 1923–1933 (eng).

[49] Karl M. Kuntzelman, Jacob M. Williams, Phui Cheng Lim, Ashok Samal, Prahalada K. Rao, and Matthew R. Johnson, *Deep-Learning-Based Multivariate Pattern Analysis (dMVPA): A Tutorial and a Toolbox*, Frontiers in Human Neuroscience **15** (2021), 638052.

[50] Fast Forward Labs, *Supercharging Classification - The Value of Multi-task Learnin*, June 2018.

[51] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, *Gradient-based learning applied to document recognition*, Proceedings of the IEEE **86** (1998), no. 11, 2278–2324.

[52] Lee, Jung, and Loui, *Attention Modulates Electrophysiological Responses to Simultaneous Music and Language Syntax Processing*, Brain Sciences **9** (2019), no. 11, 305 (en).

[53] Yune-Sang Lee, Petr Janata, Carlton Frost, Michael Hanke, and Richard Granger, *Investigation of melodic contour processing in the brain using multivariate pattern-based fMRI*, NeuroImage **57** (2011), no. 1, 293–300 (en).

[54] Kaiming Li, Lei Guo, Jingxin Nie, Gang Li, and Tianming Liu, *Review of methods for functional brain connectivity detection using fMRI*, Computerized Medical Imaging and Graphics: The Official Journal of the Computerized Medical Imaging Society **33** (2009), no. 2, 131–139 (eng).

[55] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang, *VisualBERT: A Simple and Performant Baseline for Vision and Language*, August 2019.

[56] Shiyang Li, Xiaoyong Jin, Yao Xuan, Xiyou Zhou, Wenhu Chen, Yu-Xiang Wang, and Xifeng Yan, *Enhancing the Locality and Breaking the Memory Bottleneck of Transformer on Time Series Forecasting*, January 2020.

[57] Wei Li, Xuefeng Lin, and Xi Chen, *Detecting Alzheimer's disease Based on 4D fMRI: An exploration under deep learning framework*, Neurocomputing **388** (2020), 280–287 (en).

[58] Xiaoxiao Li, Yuan Zhou, Nicha Dvornek, Muhan Zhang, Siyuan Gao, Juntang Zhuang, Dustin Scheinost, Lawrence H. Staib, Pamela Ventola, and James S. Duncan, *BrainGNN: Interpretable Brain Graph Neural Network for fMRI Analysis*, Medical Image Analysis **74** (2021), 102233 (eng).

[59] Martin A. Lindquist, *The Statistical Analysis of fMRI Data*, Statistical Science **23** (2008), no. 4.

[60] Psyche Loui, David L. Wessel, and Carla L. Hudson Kam, *Humans Rapidly Learn Grammatical Structure in a New Musical Scale*, Music Perception **27** (2010), no. 5, 377–388 (en).

[61] Yizhuo Lu, Changde Du, Dianpeng Wang, and Huiguang He, *MindDiffuser: Controlled Image Reconstruction from Human Brain Activity with Semantic and Structural Diffusion*, (2023), Publisher: arXiv Version Number: 1.

[62] Itzik Malkiel, Gony Rosenman, Lior Wolf, and Talma Hendler, *Self-Supervised Transformers for fMRI representation*, August 2022.

[63] Guilhem Marion, Giovanni M. Di Liberto, and Shihab A. Shamma, *The Music of Silence. Part I: Responses to Musical Imagery Encode Melodic Expectations and Acoustics*, The Journal of Neuroscience (2021), JN–RM–0183–21 (en).

[64] Lloyd May, Andrea R. Halpern, Sean D. Paulsen, and Michael A. Casey, *Imagined Musical Scale Relationships Decoded from Auditory Cortex*, Journal of Cognitive Neuroscience **34** (2022), no. 8, 1326–1339, _eprint: https://direct.mit.edu/jocn/article-pdf/34/8/1326/2033220/jocn_a_01858.pdf.

[65] Paul McCarthy, *FSLeyes*, August 2022, Version Number: 1.5.0.

[66] Martin Monti, *Statistical Analysis of fMRI Time-Series: A Critical Review of the GLM Approach*, Frontiers in Human Neuroscience **5** (2011).

[67] Tomoya Nakai, Naoko Koide-Majima, and Shinji Nishimoto, *Correspondence of categorical and feature-based representations of music in the human brain*, Brain and Behavior **11** (2021), no. 1, e01936, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/brb3.1936.

[68] Yun Nan, Thomas R. Knösche, Stefan Zysset, and Angela D. Friederici, *Cross-cultural music phrase processing: An fMRI study*, Human Brain Mapping **29** (2008), no. 3, 312–328 (en).

[69] Andrea Nani, Jordi Manuello, Lorenzo Mancuso, Donato Liloia, Tommaso Costa, and Franco Cauda, *The Neural Correlates of Consciousness and Attention: Two Sister Processes of the Brain*, Frontiers in Neuroscience **13** (2019), 1169.

[70] Sam Nguyen, Brenda Ng, Alan D. Kaplan, and Priyadip Ray, *Attend and Decode: 4D fMRI Task State Decoding Using Attention Models*, January 2021.

[71] Shinji Nishimoto, An T. Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L. Gallant, *Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies*, Current Biology **21** (2011), no. 19, 1641–1646 (en).

[72] Chen Niu, Alexander D. Cohen, Xin Wen, Ziyi Chen, Pan Lin, Xin Liu, Bjoern H. Menze, Benedikt Wiestler, Yang Wang, and Ming Zhang, *Modeling motor task activation from resting-state fMRI using machine learning in individual subjects*, Brain Imaging and Behavior **15** (2021), no. 1, 122–132 (en).

[73] Kenneth A. Norman, Sean M. Polyn, Greg J. Detre, and James V. Haxby, *Beyond mind-reading: multi-voxel pattern analysis of fMRI data*, Trends in Cognitive Sciences **10** (2006), no. 9, 424–430 (eng).

[74] S Ogawa, D W Tank, R Menon, J M Ellermann, S G Kim, H Merkle, and K Ugurbil, *Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging.*, Proceedings of the National Academy of Sciences **89** (1992), no. 13, 5951–5955 (en).

[75] J.M. Ollinger, G.L. Shulman, and M. Corbetta, *Separating Processes within a Trial in Event-Related Functional MRI*, NeuroImage **13** (2001), no. 1, 210–217 (en).

[76] Diana Omigie, Marcus Pearce, Katia Lehongre, Dominique Hasboun, Vincent Navarro, Claude Adam, and Severine Samson, *Intracranial Recordings and*

Computational Modeling of Music Reveal the Time Course of Prediction Error Signaling in Frontal and Temporal Cortices, Journal of Cognitive Neuroscience **31** (2019), no. 6, 855–873 (en).

[77] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu, *Neural Discrete Representation Learning*, (2017), Publisher: arXiv Version Number: 2.

[78] Janne M. Papma, Marion Smits, Marius de Groot, Francesco U. Mattace Raso, Aad van der Lugt, Henri A. Vrooman, Wiro J. Niessen, Peter J. Koudstaal, John C. van Swieten, Frederik M. van der Veen, and Niels D. Prins, *The effect of hippocampal function, volume and connectivity on posterior cingulate cortex functioning during episodic memory fMRI in mild cognitive impairment*, European Radiology **27** (2017), no. 9, 3716–3724 (eng).

[79] Sean Paulsen, Lloyd May, and Michael Casey, *Decoding Imagined Auditory Pitch Phenomena with an Autoencoder Based Temporal Convolutional Architecture*, BRAININFO (Nice, France), IARIA, July 2021.

[80] W. D. Penny, K. J. Friston, J. T. Ashburner, S. J. Kiebel, and T. E. Nichols, *Statistical parametric mapping: the analysis of functional brain images*, Elsevier, 2011.

[81] R. Penrose, *On best approximate solutions of linear matrix equations*, Mathematical Proceedings of the Cambridge Philosophical Society **52** (1956), no. 1, 17–19 (en).

[82] Francisco Pereira, Tom Mitchell, and Matthew Botvinick, *Machine learning classifiers and fMRI: A tutorial overview*, NeuroImage **45** (2009), no. 1, S199–S209 (en).

[83] Fernando J. Pineda, *Generalization of back-propagation to recurrent neural networks*, Physical Review Letters **59** (1987), no. 19, 2229–2232 (en).

[84] Russell A. Poldrack, Deanna M. Barch, Jason P. Mitchell, Tor D. Wager, Anthony D. Wagner, Joseph T. Devlin, Chad Cumba, Oluwasanmi Koyejo, and Michael P. Milham, *Toward open sharing of task-based fMRI data: the OpenfMRI project*, Frontiers in Neuroinformatics **7** (2013).

[85] Nina Politimou, Pedro Douglass-Kirk, Marcus Pearce, Lauren Stewart, and Fabia Franco, *Melodic expectations in 5- and 6-year-old children*, Journal of Experimental Child Psychology **203** (2021), 105020 (en).

[86] Marina Pominova, Alexey Artemov, Maksim Sharaev, Ekaterina Kondrateva, Alexander Bernstein, and Evgeny Burnaev, *Voxelwise 3D Convolutional and Recurrent Neural Networks for Epilepsy and Depression Diagnostics from Structural and Functional MRI Data*, 2018 IEEE International Conference on Data Mining Workshops (ICDMW) (Singapore, Singapore), IEEE, November 2018, pp. 299–307.

[87] Jonathan D. Power, Kelly A. Barnes, Abraham Z. Snyder, Bradley L. Schlaggar, and Steven E. Petersen, *Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion*, NeuroImage **59** (2012), no. 3, 2142–2154 (en).

[88] Jonathan D. Power, Mark Plitt, Timothy O. Laumann, and Alex Martin, *Sources and implications of whole-brain fMRI signals in humans*, NeuroImage **146** (2017), 609–625 (en).

[89] Alec Radford and Karthik Narasimhan, *Improving Language Understanding by Generative Pre-Training*, 2018.

[90] Rajeev D. S. Raizada, Feng-Ming Tsao, Huei-Mei Liu, and Patricia K. Kuhl, *Quantifying the Adequacy of Neural Representations for a Cross-Language Phonetic Discrimination Task: Prediction of Individual Differences*, Cerebral Cortex **20** (2010), no. 1, 1–12 (en).

[91] J. C. Rajapakse, F. Kruggel, J. M. Maisog, and D. Y. von Cramon, *Modeling hemodynamic response for analysis of functional MRI time-series*, Human Brain Mapping **6** (1998), no. 4, 283–300 (eng).

[92] Ali Razavi, Aaron van den Oord, and Oriol Vinyals, *Generating Diverse High-Fidelity Images with VQ-VAE-2*, (2019), Publisher: arXiv Version Number: 1.

[93] Baxter P. Rogers, Victoria L. Morgan, Allen T. Newton, and John C. Gore, *Assessing functional connectivity in the human brain by fMRI*, Magnetic Resonance Imaging **25** (2007), no. 10, 1347–1357 (eng).

[94] Bruce R. Rosen, Randy L. Buckner, and Anders M. Dale, *Event-related functional MRI: Past, present, and future*, Proceedings of the National Academy of Sciences **95** (1998), no. 3, 773–780 (en).

[95] Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf, *Transfer Learning in Natural Language Processing*, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials (Minneapolis, Minnesota), Association for Computational Linguistics, June 2019, pp. 15–18.

[96] Valorie N. Salimpoor, David H. Zald, Robert J. Zatorre, Alain Dagher, and Anthony Randal McIntosh, *Predictions and the brain: how musical sounds become rewarding*, Trends in Cognitive Sciences **19** (2015), no. 2, 86–91.

[97] Sarah A. Sauvé, Aminah Sayed, Roger T. Dean, and Marcus T. Pearce, *Effects of pitch and timing expectancy on musical emotion.*, Psychomusicology: Music, Mind, and Brain **28** (2018), no. 1, 17–39 (en).

[98] Alexander Schaefer, Ru Kong, Evan M Gordon, Timothy O Laumann, Xi-Nian Zuo, Avram J Holmes, Simon B Eickhoff, and B T Thomas Yeo, *Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI*, Cerebral Cortex **28** (2018), no. 9, 3095–3114 (en).

[99] Per Sidén, *Scalable Bayesian spatial analysis with Gaussian Markov random fields*, September 2020.

[100] Aleksi J Sihvonen, Teppo Särkämö, Vera Leo, Mari Tervaniemi, Eckart Altenmüller, and Seppo Soinila, *Music-based interventions in neurological rehabilitation*, The Lancet Neurology **16** (2017), no. 8, 648–660 (en).

[101] Olivier Simon, Ferath Kherif, Guillaume Flandin, Jean-Baptiste Poline, Denis Rivière, Jean-François Mangin, Denis Le Bihan, and Stanislas Dehaene, *Automatized clustering and functional geometry of human parietofrontal networks for language, space, and number*, NeuroImage **23** (2004), no. 3, 1192–1202 (en).

[102] Xiaomu Song and Nan-kuei Chen, *A SVM-based quantitative fMRI method for resting-state functional network detection*, Magnetic Resonance Imaging **32** (2014), no. 7, 819–831 (eng).

[103] Noël Staeren, Hanna Renvall, Federico De Martino, Rainer Goebel, and Elia Formisano, *Sound Categories Are Represented as Distributed Patterns in the Human Auditory Cortex*, Current Biology **19** (2009), no. 6, 498–502 (en).

[104] S.C. Strother, *Evaluating fMRI preprocessing pipelines*, IEEE Engineering in Medicine and Biology Magazine **25** (2006), no. 2, 27–41.

[105] Heung-Il Suk, Chong-Yaw Wee, Seong-Whan Lee, and Dinggang Shen, *State-space model with deep learning for functional dynamics estimation in resting-state fMRI*, NeuroImage **129** (2016), 292–307 (eng).

[106] Shoji Tanaka and Eiji Kirino, *Functional Connectivity of the Dorsal Striatum in Female Musicians*, Frontiers in Human Neuroscience **10** (2016).

[107] Michael H. Thaut and Gerald C. McIntosh, *Neurologic Music Therapy in Stroke Rehabilitation*, Current Physical Medicine and Rehabilitation Reports **2** (2014), no. 2, 106–113 (en).

[108] Armin W. Thomas, Hauke R. Heekeren, Klaus-Robert Müller, and Wojciech Samek, *Analyzing Neuroimaging Data Through Recurrent Deep Learning Models*, April 2019.

[109] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang, *VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training*, (2022), Publisher: arXiv Version Number: 3.

[110] G. Tzanetakis and P. Cook, *Musical genre classification of audio signals*, IEEE Transactions on Speech and Audio Processing **10** (2002), no. 5, 293–302.

[111] Koene R. A. Van Dijk, Trey Hedden, Archana Venkataraman, Karleyton C. Evans, Sara W. Lazar, and Randy L. Buckner, *Intrinsic functional connectivity as a tool for human connectomics: theory, properties, and optimization*, Journal of Neurophysiology **103** (2010), no. 1, 297–321 (eng).

[112] David C. Van Essen, Stephen M. Smith, Deanna M. Barch, Timothy E. J. Behrens, Essa Yacoub, Kamil Ugurbil, and WU-Minn HCP Consortium, *The WU-Minn Human Connectome Project: an overview*, NeuroImage **80** (2013), 62–79 (eng).

[113] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, *Attention Is All You Need*, December 2017.

[114] Archana Venkataraman, Koene R. A. Van Dijk, Randy L. Buckner, and Polina Golland, *EXPLORING FUNCTIONAL CONNECTIVITY IN FMRI VIA CLUSTERING*, Proceedings of the ... IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP (Conference) **2009** (2009), 441–444 (eng).

[115] Archana Venkataraman, Daniel Y.-J. Yang, Kevin A. Pelphrey, and James S. Duncan, *Bayesian Community Detection in the Space of Group-Level Functional Differences*, IEEE Transactions on Medical Imaging **35** (2016), no. 8, 1866–1882.

[116] Sandra Vieira, Walter H.L. Pinaya, and Andrea Mechelli, *Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications*, Neuroscience & Biobehavioral Reviews **74** (2017), 58–75 (en).

[117] Defeng Wang, Lin Shi, Daniel S. Yeung, Pheng-Ann Heng, Tien-Tsin Wong, and Eric C. C. Tsang, *Support vector clustering for brain activation detection*, Medical image computing and computer-assisted intervention: MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention **8** (2005), no. Pt 1, 572–579 (eng).

[118] Xiaoxiao Wang, Xiao Liang, Zhoufan Jiang, Benedictor Alexander Nguchu, Yawen Zhou, Yanming Wang, Huijuan Wang, Yu Li, Yuying Zhu, Feng Wu, Jia-Hong Gao, and Benching Qiu, *Decoding and mapping task states of the human brain via deep learning*, Human Brain Mapping **41** (2020), no. 6, 1505–1519.

[119] Ze Wang, Anna R. Childress, Jiongjiong Wang, and John A. Detre, *Support vector machine learning-based fMRI data group analysis*, NeuroImage **36** (2007), no. 4, 1139–1151 (eng).

[120] Neil D. Woodward and Carissa J. Cascio, *Resting-State Functional Connectivity in Psychiatric Disorders*, JAMA psychiatry **72** (2015), no. 8, 743–744 (eng).

[121] M. W. Woolrich, B. D. Ripley, M. Brady, and S. M. Smith, *Temporal autocorrelation in univariate linear modeling of FMRI data*, NeuroImage **14** (2001), no. 6, 1370–1386 (eng).

[122] K. J. Worsley, A. C. Evans, S. Marrett, and P. Neelin, *A Three-Dimensional Statistical Analysis for CBF Activation Studies in Human Brain*, Journal of Cerebral Blood Flow & Metabolism **12** (1992), no. 6, 900–918 (en).

[123] K.J. Worsley and K.J. Friston, *Analysis of fMRI Time-Series Revisited—Again*, NeuroImage **2** (1995), no. 3, 173–181 (en).

[124] Cedric Huchuan Xia, Zongming Ma, Rastko Ciric, Shi Gu, Richard F. Betzel, Antonia N. Kaczkurkin, Monica E. Calkins, Philip A. Cook, Angel García de la Garza, Simon N. Vandekar, Zaixu Cui, Tyler M. Moore, David R. Roalf, Kosha Ruparel, Daniel H. Wolf, Christos Davatzikos, Ruben C. Gur, Raquel E. Gur, Russell T. Shinohara, Danielle S. Bassett, and Theodore D. Satterthwaite, *Linked dimensions of psychopathology and connectivity in functional brain networks*, Nature Communications **9** (2018), no. 1, 3003 (eng).

[125] Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo, *Performance-optimized hierarchical models predict neural responses in higher visual cortex*, Proceedings of the National Academy of Sciences **111** (2014), no. 23, 8619–8624 (en).

[126] B. T. Thomas Yeo, Fenna M. Krienen, Jorge Sepulcre, Mert R. Sabuncu, Danial Lashkari, Marisa Hollinshead, Joshua L. Roffman, Jordan W. Smoller, Lilla Zöllei, Jonathan R. Polimeni, Bruce Fischl, Hesheng Liu, and Randy L. Buckner, *The organization of the human cerebral cortex estimated by intrinsic functional connectivity*, Journal of Neurophysiology **106** (2011), no. 3, 1125–1165 (eng).

[127] Robert J. Zatorre, *Cerebral Correlates of Human Auditory Processing*, Acoustical Signal Processing in the Central Auditory System (Josef Syka, ed.), Springer US, Boston, MA, 1997, pp. 453–468 (en).

[128] Robert J. Zatorre and Valorie N. Salimpoor, *From perception to pleasure: Music and its neural substrates*, Proceedings of the National Academy of Sciences **110** (2013), no. supplement_2, 10430–10437 (en).

[129] Ali Zeineddine, Amor Nafkha, Stéphane Paquelet, Christophe Moy, and Pierre Yves Jezequel, *Comprehensive Survey of FIR-Based Sample Rate Conversion*, Journal of Signal Processing Systems **93** (2021), no. 1, 113–125 (en).

[130] Xiaoyan Zhan and Rongjun Yu, *A Window into the Brain: Advances in Psychiatric fMRI*, BioMed Research International **2015** (2015), 542467 (eng).

[131] Chongyue Zhao, Hongming Li, Zhicheng Jiao, Tianming Du, and Yong Fan, *A 3D Convolutional Encapsulated Long Short-Term Memory (3DConv-LSTM) Model for Denoising fMRI Data*, Medical image computing and computer-assisted intervention: MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention **12267** (2020), 479–488 (eng).

[132] Liang Zou, Jiannan Zheng, Chunyan Miao, Martin J. Mckeown, and Z. Jane Wang, *3D CNN Based Automatic Diagnosis of Attention Deficit Hyperactivity*

Disorder Using Functional and Structural MRI, IEEE Access **5** (2017), 23626–23636.