Unsupervised Representation Learning of Brain Activity via Bridging Voxel Activity and Functional Connectivity

Anonymous Author(s)
Affiliation
Address
email

Abstract

Effective brain representation learning is a key step toward revealing the understanding of cognitive processes and unlocking detecting and potential therapeutic interventions for neurological diseases/disorders. Existing studies have focused on either (1) voxel-level activity, where only a single beta weight for each voxel (i.e., aggregation of voxel activity over a time window) is considered, missing their temporal dynamics, or (2) functional connectivity of the brain in the level of region of interests, missing voxel-level activities. In this paper, we bridge this gap and design BrainMixer, an unsupervised learning framework that effectively utilizes both functional connectivity and associated time series of voxels to learn voxel-level representation in an unsupervised manner. BrainMixer employs two simple yet effective MLP-based encoders to simultaneously learn the dynamics of voxel-level signals and their functional correlations. To encode voxel activity, BrainMixer fuses information across both time and voxel dimensions via a dynamic self-attention mechanism. To learn the structure of the functional connectivity graph, BrainMixer presents a temporal graph patching and encodes each patch by combining its nodes' features via a new adaptive temporal graph pooling. Our experiments show that BrainMixer attains outstanding performance and outperforms 13 baselines in different downstream tasks and experimental setups.

1 Introduction

Understanding the human brain is a long-term intriguing goal for neuroscience and recent advancements in machine learning methods have provided powerful paradigms to achieve this goal (Guo et al., 2016; Poldrack & Farah, 2015). While neuroimaging techniques, as the principal source of brain data, provide rich information about brain functions, the provided data is high-dimensional and complex in nature (Poldrack & Gorgolewski, 2014). To overcome this challenge, representation learning serves as the backbone of machine learning methods on neuroimaging data and provides a low-dimensional representation of brain components at different levels of granularity, enabling the understanding of behaviors (Schneider et al., 2023), brain functions (Yamins & DiCarlo, 2016) and/or detecting neurological diseases or disorders (Behrouz & Seltzer, 2023a; Uddin et al., 2017).

In the brain imaging literature, studies have mainly focused on two spatial scales—voxel-level and network-level—as well as two analysis approaches—multivariate pattern analysis (MVPA) and functional connectivity (Mahmoudi et al., 2012; Van Den Heuvel & Pol, 2010). The MVPA approach is often employed at the voxel-level scale and in task-based studies to associate neural activities at a very fine-grained and local level with particular cognitive functions, behaviors, or stimuli. This method has found applications in various areas, including the detection of neurological conditions (Sundermann et al., 2014; Bray et al., 2009), neurofeedback interventions (Cortese et al.,

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

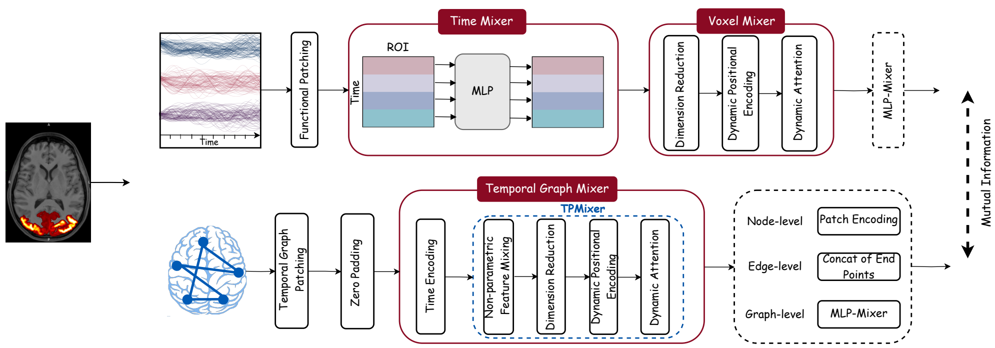

Figure 1: **Schematic of the BRAINMIXER**. BRAINMIXER consists of two main modules: (1) Voxel Activity Encoder (top), and (2) Functional Connectivity Encoder (bottom).

2021), decoding neural responses to visual stimuli (Horikawa & Kamitani, 2017), deciphering memory contents (Lee & Baker, 2016; Chadwick et al., 2012), and classifying cognitive states (Mitchell et al., 2003). The functional connectivity analysis, on the other hand, focuses on the temporal correlations or statistical dependencies between the activity of different brain regions at larger scales to assess how these areas communicate and collaborate. This method has been utilized to study various topics such as task-related network dynamics (Gonzalez-Castillo & Bandettini, 2018; Hutchison et al., 2013) and the effects of neurological disorders on brain connectivity (Greicius, 2008; Du et al., 2018).

Limitation of Previous Methods. Despite the advances in the representation learning of brain signals, existing studies suffer from a subset of five limitations: ① Study the human brain at a single scale: Most existing studies study the brain at either voxel-level or functional connectivity, while these two scales can provide complementary information to each other; e.g., although voxel-level activity provides detailed and more accurate information about brain activity, it misses the information about how different areas communicate with each other at a high level. Recently, this limitation has motivated researchers to search for new methods of integrating these two levels of analyses (Nieto-Castanon, 2022; McNorgan et al., 2020). ② Supervised setting: Learning brain activity in a supervised setting relies on a large number of clinical labels while obtaining accurate and reliable clinical labels is challenging due to its high cost (Avberšek & Repovš, 2022).③ Missing information by averaging: Most existing studies on voxel activities aggregate measured voxel activity (e.g., its blood-oxygen level dependence) over each time window to obtain a single beta weight (Roth et al., 2022; Vassena et al., 2020; Roth & Merriam, 2023). However, this approach misses the voxel activity dynamic over each task. Moreover, most studies on brain functional connectivity also aggregate closed voxels to obtain brain activity in the Region of Interest (ROI) level, missing individual voxel activities. ④ Missing the dynamics of the interactions: Some existing studies neglect the fact that the functional connectivity of the human brain dynamically changes over time, even in resting-state neuroimaging data (Calhoun et al., 2014). In task-dependent neuroimage data, subjects are asked to perform different tasks in different time windows, and the dynamics of the brain activity play an important role in understanding neurological disease/disorder (Hernandez et al., 2015). ⑤ Designed for a particular task or neuroimaging modality: Due to the different and complex clinical patterns of brain signals (da Silva, 1991), some existing methods are designed for a particular type of brain signal data (Lanciano et al., 2020; Cai et al., 2023), and there is a lack of a unified framework.

Application to Understanding Object Representation in the Brain. Understanding object representation in the brain is a key step toward revealing the basic building blocks of human visual processing (Hebart et al., 2023). Due to the hierarchical nature of human visual processing, it requires analyzing brain activity at different scales, i.e., both functional connectivity graph and voxel-level activity. However, there is a small number of studies in this area, possibly due to the lack of proper large-scale datasets. In this study, we present two large-scale graph-structured datasets, BVFC and BVFC-MEG, based on raw fMRI and MEG from THINGS (Hebart et al., 2023) dataset. BVFC (resp. BVFC-MEG) comprises 26,220 graphs (resp. 89,792 graphs) with up to 13,166 nodes (resp. 272 nodes), representing brain voxels' activity (resp. channels) in functional MRI (fMRI) (resp. magnetoencephalographic (MEG)) for human subjects when seeing natural or unrecognizable GAN-generated images. We believe this dataset can bridge graph anomaly detection and graph classification tasks to understanding object representation in the brain (see §4). See Appendix B for more details.

Contributions. To overcome the above limitations, we leverage both voxel-level activity and functional connectivity of the brain. We present BRAINMIXER, an unsupervised MLP-based brain representation learning approach that jointly learns voxel-level activity and functional connectivity. BRAINMIXER employs a novel multivariate timeseries encoder that binds information across both time and voxel dimensions. It uses a simple MLP with functional patching to fuse information across different timestamps and learns dynamic self-attention weights to fuse information across voxels based on their functionality. On the other hand, BRAINMIXER uses a novel temporal graph learning method to encode the brain functional connectivity. The graph encoder first extracts temporal patches using temporal random walks and then fuses information within each patch using the designed dynamic self-attention mechanism. We further propose an adaptive permutation invariant pooling to obtain patch encodings. Since voxel activity and functional connectivity encodings are different views of the same context, we propose an unsupervised pre-training approach to jointly learn voxel activity and functional connectivity by maximizing their mutual information. In the experimental evaluations, we provide two new large-scale graph and timeseries datasets based on THINGS (Hebart et al., 2023). Extensive experiments on six datasets show the superior performance of BRAINMIXER and the significance of each of its components in a variety of downstream tasks.

For the sake of consistency, we explain BRAINMIXER for fMRI modality; however, as it is shown in §4, it can simply be used for any other neuroimaging modalities that provide a timeseries for each part of the brain (e.g., MEG and EEG). When dealing with MEG or EEG, we can replace the term "voxel" with "channel". Supplementary materials (code and Appendix) can be found in this link.

2 Related Work

To situate our BRAINMIXER in a broader context, we briefly review machine learning models for timeseries, graphs, and neuroscience. For extensive discussion of related work see Appendix C.

Timeseries Learning. Attention mechanisms are powerful models to capture long-range dependencies and so recently, Transformer-based models have attracted much attention in time series forecasting (Zerveas et al., 2021; Li et al., 2019). Due to their quadratic time complexity, several studies aim to reduce the time and memory usage of these methods (Child et al., 2019). Another type of work uses (hyper)graph learning frameworks to learn (higher-order) patterns in timeseries (Park et al., 2009; Sawhney et al., 2021). Inspired by the recent success of MLP-MIXER (Tolstikhin et al., 2021), Li et al. (2023) and Chen et al. (2023) presented two variants of MLP-MIXER for timeseries forecasting. All these methods are different from BRAINMIXER, as ① they use static attention mechanisms, ② do not take advantage of the functionality of voxels in patching, and ③ are designed for timeseries forecasting and cannot simply be extended to various downstream tasks on the brain.

MLP-based Graphs Learning. Learning on graphs has been an active research area in recent years (Jiang et al., 2021; Veličković et al., 2018; Chamberlain et al., 2023). While most studies use message-passing frameworks to learn the local and global structure of the graph, recently, due to the success of MLP-based methods (Tolstikhin et al., 2021), MLP-based graph learning methods have attracted much attention (Hu et al., 2021; Behrouz et al., 2023). For example, Cong et al. (2023) and He et al. (2023) presented two extensions of MLP-MIXER to graph-structured data. However, all these methods are different from BRAINMIXER and specifically FC Encoder, as either ① use time-consuming graph clustering algorithms for patching, ② are static methods and cannot capture temporal properties, or ③ are attention-free and cannot capture the importance of nodes.

Graph Learning and Timeseries for Neuroscience. In recent years, several studies have analyzed functional connectivity to differentiate human brains with a neurological disease/disorder (Jie et al., 2016; Chen et al., 2011; Wee et al., 2011). With the success of graph neural networks in graph data analysis, deep learning models have been developed to predict brain diseases by studying brain network structures (Behrouz & Seltzer, 2022; Zhu et al., 2022; Cui et al., 2022). Moreover, several studies focus on brain signals (Craik et al., 2019; Shoeibi et al., 2021) to detect neurological diseases. For example, Cai et al. (2023) designed a self-supervised learning framework to detect seizures from EEG and SEEG data. However, all these methods are different from BRAINMIXER as they are designed for a particular task (e.g., brain classification), a particular neuroimaging modality (e.g., fMRI or EEG), and/or supervised settings.

3 Method: BrainMixer

In this section, we first discuss the notation we use throughout the paper. Detailed discussion about background concepts can be found in Appendix A.

Notation. We represent the neuroimaging of a human brain as $\mathcal{B} = \{\mathcal{B}^{(t)}\}_{t=1}^{T}$ where $\mathcal{B}^{(t)} = (\mathcal{V}, \mathcal{G}_F^{(t)}, \mathcal{X}^{(t)}, \mathbb{F})$ represents the neural data in time window $1 \leq t \leq T$. Here, $\mathcal{V}$ is the set of voxels, $\mathcal{G}_F^{(t)} = (\mathcal{V}, \mathcal{E}^{(t)}, \mathcal{A}^{(t)})$ is the functional connectivity graph, $\mathcal{E}^{(t)} \subseteq \mathcal{V} \times \mathcal{V}$ is the set connections between voxels, $\mathcal{A}^{(t)}$ is the correlation matrix (weighted adjacency matrix of $\mathcal{G}_F^{(t)}$), $\mathcal{X}^{(t)} \in \mathbb{R}^{|\mathcal{V}| \times \tilde{T}}$ is a multivariate timeseries of voxels activities, and $\mathbb{F}$ is the set of functional systems in the brain (Schaefer et al., 2018) in time window t.

3.1 Voxel Activity Encoder

The main goal of this module is to learn the time series of the voxel-level activity. However, the activities of voxels are not disjoint; for example, an increase in fusiform face area (FFA) activity might be associated with a rise in V1 activity. Accordingly, effectively learning their dynamics patterns requires both capturing cross-voxel and within-voxel time series information. The vanilla MLP-Mixer (Tolstikhin et al., 2021) can be used to bind information across both of these dimensions, but the human brain has unique traits that make directly applying MLP-Mixer insufficient/impractical. ① There does not exist in general a canonical grid of the brain to encode voxel activities, which makes patch extraction challenging. ② Contrary to images that can be divided into patches of the same size, the partitioning of voxels might not be all the same size due to the complex brain topology. ③ MLP-Mixer employs a fixed static mixing matrix for binding patches, while in the brain the functionality of each token is important and a different set of tokens should be mixed differently based on their connections and functionality. To address these challenges, the *VA Encoder* employs two submodules, *time-mixer* and *voxel-mixer* with dynamic mixing matrix, to fuse information across both time and voxel dimensions, respectively.

The human brain is comprised of functional systems (FS) (Schaefer et al., 2018), which are groups of voxels that perform similar functions (Smith et al., 2013). We take advantage of this hierarchical structure and patch voxels based on their functionality. However, the main challenge is that the sizes of the patches (set of voxels with similar functionality) are different. To this end, inspired by the inference of ViT models (Dosovitskiy et al., 2021), we linearly interpolate patches with smaller sizes.

Functional Patching. Let K be the number of voxels and $\mathbf{X} \in \mathbb{R}^{K \times (T \times \tilde{T})}$ represents the time series of voxels activities over all time windows. We split $\mathbf{X}$ to spatio-temporal patches $\mathbf{X}_i$ with size $|f_i| \times t_p$, where $f_i \in \mathbb{F}$ is a functional system (Schaefer et al., 2018), and t_p is the temporal-dimension length. To address the challenge of patches with different sizes, we use $\textsc{Interpolate}(.)$ to linearly interpolate patches to the same size N_p: i.e., $\tilde{\mathbf{X}}_i = \textsc{Interpolate}(\mathbf{X}_i)$, where $\tilde{\mathbf{X}}_i \in \mathbb{R}^{N_p \times t_p}$.

Voxel-Mixer. Since the effect of each task (e.g., in task-based fMRI) on brain activity as well as the time it lasts varies (Yang et al., 2023a), for different tasks, we might need to emphasize more on a subset of voxels. To this end, to bind information across voxels, we use a dynamic attention mechanism that uses a learnable dynamic mixing matrix $\mathbf{P}_i$, learning to mix a set of input voxels based on their functionality. While using different learnable matrices for mixing voxels activity provides a more powerful architecture, its main challenge is a large number of parameters. To mitigate this challenge, we first reduce the dimensions of $\tilde{\mathbf{X}}$, split it into a set of segments, denoted as S, and then combine the transformed matrices. Given a segment $s \in S$ we have:

$$\hat{\mathbf{X}}^{(t)^{(s)}} = \tilde{\mathbf{X}}^{(t)} \, \mathbf{W}_{\text{segment}}^{(s)}, \qquad (\textit{Dimension Reduction})$$

$$\mathbf{P}_i^{(s)} = \textsc{Softmax}\left(\textsc{Flat}\left(\hat{\mathbf{X}}^{(t)^{(s)}}\right) \mathbf{W}_{\text{flat}}^{(s)^{(i)}}\right), \qquad (\textit{Learning Dynamic Mixer})$$

$$\mathbf{X}_{\text{PE}}^{(t)} = \left[\Big\|_{s \in S} \mathbf{P}^{(s)} \tilde{\mathbf{X}}^{(t)^{(s)}} \right] \mathbf{W}_{\text{PE}}, \qquad (\textit{Dynamic Positional Encoding})$$

$$\mathbf{H}_{\text{Voxel}}^{(t)} = \texttt{Norm}\left(\tilde{\mathbf{X}}^{(t)}\right) + \textsc{Sigmoid}\left(\frac{\mathbf{X}_{\text{PE}}^{(t)} \, \mathbf{X}_{\text{PE}}^{(t)^{\top}}}{\sqrt{\tilde{T}}}\right) \mathbf{X}_{\text{PE}}^{(t)}, \qquad (\textit{Dynamic Self-Attention})$$

where $\mathbf{W}^{(s)}_{\text{segment}} \in \mathbb{R}^{\tilde{T}\times d}$, $\mathbf{W}^{(s)^{(i)}}_{\text{flat}} \in \mathbb{R}^{(K\times d)\times K}$, $\mathbf{W}_{\text{PE}} \in \mathbb{R}^{\tilde{T}\times\tilde{T}}$ are learnable parameters, $\|$ is concatenation, and $\textsc{Sigmoid}(.)$ is row-wise sigmoid normalization. Note that for different segments we use different dimensionality reduction matrices to reinforce the power of the Voxel Mixing.

Time Mixer. We first fuse information in the time dimension by using the Time Mixer submodule. To this end, the Time Mixer employs a 2-layer MLP with layer-normalization (Ba et al., 2016):

$$\mathbf{H}^{(t)}_{\textit{Time}} = \mathbf{H}^{(t)}_{\text{Voxel}} + \left(\sigma\left(\texttt{LayerNorm}\left(\mathbf{H}^{(t)}_{\text{Voxel}}\right)\mathbf{W}^{(1)}_{\text{Time}}\right)\mathbf{W}^{(2)}_{\text{Time}}\right), \tag{1}$$

where $\mathbf{W}^{(1)}_{\text{Time}}$ and $\mathbf{W}^{(1)}_{\text{Time}}$ are learnable matrices, $\sigma(.)$ is an activation function (we use GeLU (Hendrycks & Gimpel, 2020)), and `LayerNorm` is layer normalization (Ba et al., 2016).

3.2 Functional Connectivity Graph Encoder

To encode the functional connectivity graph, we design an MLP-based architecture that learns both the structural and temporal properties of the graph. Inspired by the recent success of all-MLP architecture in graphs (Cong et al., 2023), we extend MLP-Mixer to temporal graphs. We first define patches in temporal graphs. While patches in images, videos, and multivariate timeseries can simply be non-overlapping regular grids, patches in graphs are overlapping non-grid structures, which makes the patching extraction challenging. He et al. (2023) suggest using graph partitioning algorithms to extract graph patches; however, these partitioning algorithms ① only consider structural properties, missing the temporal dependencies, and ② can be time-consuming, limiting the scalability to dense graphs like brain functional connectome. To this end, we propose a temporal-patch extraction algorithm such that nodes (voxels) in each patch share similar temporal and structural properties.

Temporal Patching. To extract temporal patches from the graph, we use a biased temporal random walk that walks over both nodes (voxels) and timestamps. Given a functional connectivity graph $\mathcal{G}_F = \{\mathcal{G}^{(t)}_F\}_{t=1}^{T}$, we sample M walks with length $m+1$ started from node (voxel) $v_0 \in \mathcal{V}$ like: $\mathcal{W}alk : (v_0, t_0) \rightarrow (v_1, t_1) \rightarrow \cdots \rightarrow (v_m, t_m)$, such that $(v_{i-1}, v_i) \in \mathcal{E}^{(t_i)}$, and $t_0 \geq t_1 \geq t_2 \geq \cdots \geq t_m$. Note that, contrary to some previous temporal random walks (Wang et al., 2021; Behrouz et al., 2023), we allow the walker to walk in the same timestamp at each step. While backtracking over time, we aim to capture temporal information and extract the dynamics of voxels' activity over *related* timestamps. Previous studies show that doing a task can affect brain activity even after 2 minuetes (Yang et al., 2023a). To this end, since more recent connections can be more informative, we use a biased sampling procedure. Let v_p be the previously sampled node, we use hyperparameters $\theta, \theta_0 \geq 0$ to sample a node v with probability proportional to $\exp\left(\theta(t - t_p + \theta_0)\right)$, where t and t_p are the timestamps that $(v_p, v) \in \mathcal{E}^{(t)}$ and the timestamp of the previous sample, respectively. In this sampling procedure, smaller (resp. larger) θ means less (resp. more) emphasis on recent timestamps. Each walk started from v can be seen as a temporal subgraph, and so we let ρ_v be the union of all these subgraphs (walks started from v). We treat each of ρ_v as a temporal patch.

Temporal Pooling Mixer. Given the temporal graph patches that we extracted above, we need to encode each patch to obtain patch encodings (we later use these patch encodings as their corresponding voxel's encodings). While simple poolings (e.g., $\textsc{Sum}(.)$) are shown to miss information (Behrouz et al., 2023), more complicated pooling functions consider a static pooling rule. However, as discussed above, the effect of performing a task on the neuroimaging data might last for a period of time and the pooling rule might change over time. To this end, we design a temporal pooling, $\text{TPM}\textsc{ixer}(.)$, that dynamically pools a set of voxels in a patch based on their timestamps.

Given a patch $\rho_{v_0} = \{v_0, v_1, \ldots, v_k\}$, for each voxel we consider the correlation of its activity with other voxels' as its preliminary feature vector. That is, for each voxel v, we consider its feature vector in the time window t as $\mathcal{A}^{(t)}_v$, the v's corresponding row in $\mathcal{A}^{(t)}$. We abuse the notation and use $\mathcal{A}^{(t)}_{\rho_v}$ to refer to the set of $\mathcal{A}^{(t)}$'s rows corresponding to ρ_v. Since patch sizes are different, we zero pad $\mathcal{A}^{(t)}_{\rho_v}$ matrices to a fixed size. Note that this zero padding is important to capture the size of each voxel neighborhood. The voxel with more zero-padded dimensions in its patch has less correlation with others. To capture both cross-feature and cross-voxel dependencies, we can use the same architecture as the Time Mixer and Voxel-Mixer. However, the main drawback of this approach is that a pooling function is expected to be permutation invariant while the Voxel Mixer phase is permutation variant. To address this challenge, we fuse information across features in a non-parametric manner as follows:

$$\mathbf{H}_{\text{F}}^{(t)} = \mathcal{A}_{\rho_v}^{(t)} + \sigma\left(\texttt{Softmax}\left(\texttt{LayerNorm}\left(\mathcal{A}_{\rho_v}^{(t)}\right)^{\top}\right)\right)^{\top}, \tag{2}$$

where $\sigma(.)$ is an activation function and $\texttt{Softmax}(.)$ is used to normalize across features to bind and fuse feature-wise information in a non-parametric manner, avoiding permutation variant operations in the Time Mixer. To dynamically fuse information across voxels, we use the same idea as dynamic self-attention in §3.1 and learn dynamic matrices $\mathbf{P}_{\text{Pool}_i}$; let d_{patch} be the patch size:

$$\mathbf{P}_{\text{Pool}_i} = \textsc{Softmax}\left(\textsc{Flat}\left(\mathbf{H}_{\text{F}}^{(t)}\right)\mathbf{W}_{\text{Pool}}^{(i)}\right) \tag{3}$$

$$\mathbf{h}_{\rho_v} = \textsc{Mean}\left(\texttt{Norm}(\mathbf{H}_{\text{F}}^{(t)}) + \mathbf{H}_{\text{PE}}^{(t)}\ \textsc{Softmax}\left(\frac{\mathbf{H}_{\text{PE}}^{(t)\top}\ \mathbf{H}_{\text{PE}}^{(t)}}{\sqrt{d_{\text{patch}}}}\right)\right), \tag{4}$$

where $\mathbf{H}_{\text{PE}}^{(t)} = \mathbf{H}_{\text{F}}^{(t)}\mathbf{P}_{\text{Pool}}$ is the transformation of $\mathbf{H}_{\text{F}}^{(t)}$ by dynamic matrix $\mathbf{P}_{\text{Pool}}$.

Theorem 1. TPMIXER *is permutation invariant and a universal approximator of multisets.*

Time Encoding. To distinguish different timestamps in the functional connectivity graph, we use a non-learnable time encoding module proposed by Cong et al. (2023). This encoding approach helps reduce the number of parameters, and also it has been shown to be more stable and generalizable (Cong et al., 2023). Given hyperparameters α, β, and d, we use feature vector $\boldsymbol{\omega} = \{\alpha^{-i/\beta}\}_{i=0}^{d-1}$ to encode each timestamp t using $\cos(\boldsymbol{\omega}t)$ function. Therefore, we obtain the time encoding as $\boldsymbol{\eta}_t = \cos(\boldsymbol{\omega}t)$.

Voxel-, Edge-, and Graph-level Encodings. Depending on the downstream task, we might obtain voxel-, edge-, or graph-level encodings. For each voxel $v \in \mathcal{V}$, we let $\mathcal{E}^{(t)}[\rho_v]$ be the set of connections in the patch of v. To obtain the voxel-level encoding of each voxel v, $\boldsymbol{\psi}_v$, we use patch encoding and concatenate it with all the weighted mean of timestamp encodings; i.e., $\boldsymbol{\psi}_v^t = \text{MLP}([\mathbf{h}_{\rho_v} \| \mathcal{T}_v])$, where $\mathcal{T}_v = \frac{\sum_{t_0=1}^{t}\mathcal{E}^{(t_0)}[\rho_v]\boldsymbol{\eta}_{t_0}}{\sum_{t_0=1}^{t}\mathcal{E}^{(t_0)}[\rho_v]}$. For a connection $e = (u,v) \in \mathcal{E}^{(t)}$, we obtain its encoding by concatenating its endpoints and its timestamp encodings; i.e., $\boldsymbol{\zeta}_{(u,v)}^{(t)} = \text{MLP}\left([\boldsymbol{\psi}_u^t, \boldsymbol{\psi}_v^t, \boldsymbol{\eta}_t]\right)$. Finally, to obtain the graph level encoding, we use vanilla MLP-MIXER (Tolstikhin et al., 2021) on patch encodings; let $\boldsymbol{\Psi}^{(t)}$ be the matrix whose rows are $\boldsymbol{\psi}_v^{(t)}$:

$$\boldsymbol{\Psi}_{\text{token}}^{(t)} = \boldsymbol{\Psi}^{(t)} + \mathbf{W}_{\text{token}}^{(2)}\sigma\left(\mathbf{W}_{\text{token}}^{(1)}\texttt{LayerNorm}\left(\boldsymbol{\Psi}^{(t)}\right)\right), \tag{5}$$

$$\textsc{Enc}(\mathcal{G}_F^{(t)}) = \textsc{Mean}\left(\boldsymbol{\Psi}_{\text{token}}^{(t)} + \sigma\left(\texttt{LayerNorm}\left(\boldsymbol{\Psi}_{\text{token}}^{(t)}\right)\mathbf{W}_{\text{channel}}^{(1)}\right)\mathbf{W}_{\text{channel}}^{(2)}\right). \tag{6}$$

3.3 Self-supervised Pre-training

In §3.1 and §3.2 we obtained the encodings of the same contexts, from different perspectives. In this section, inspired by (Hjelm et al., 2019; Bachman et al., 2019), we use the mutual information of these two perspectives from the same context, to learn voxel- and brain-level encodings in a self-supervised manner. To this end, let $\boldsymbol{\Psi}$ be the voxel-level encodings obtained from functional connectome, $\mathbf{Z}_{\mathbf{F}}^{(t)} = \textsc{Enc}(\mathcal{G}_F^{(t)})$ be the global encoding (brain-level) of the functional connectome, $\mathbf{H}_{\text{Voxel}}^{(t)}$ be the voxel activity encodings from the brain activity timeseries, and $\mathbf{Z}_{\mathbf{V}}^{(t)}$ be the global encoding (brain-level) of the voxel activity timeseries, we aim to maximize $I(\mathbf{Z}_{\mathbf{F}}^{(t)}, \boldsymbol{\psi}_{v,i}^{(t)}) + I(\mathbf{Z}_{\mathbf{V}}^{(t)}, (\mathbf{H}_{\text{Voxel}}^{(t)})_{v,j})$ for all $v \in \mathcal{V}$ and possible i, j. Following previous studies (Bachman et al., 2019), we use Noise-Contrastive Estimation (NCE) (Gutmann & Hyvärinen, 2010) and minimize the following loss function:

$$\mathbb{E}_{(\mathbf{Z}_{\mathbf{F}}^{(t)}, \boldsymbol{\psi}_{v,i}^{(t)})}\left[\mathbb{E}_{\mathcal{N}}\left[\mathcal{L}_{\Phi}(\mathbf{Z}_{\mathbf{F}}^{(t)}, \boldsymbol{\psi}_{v,i}^{(t)}, \mathcal{N})\right]\right] + \mathbb{E}_{(\mathbf{Z}_{\mathbf{V}}^{(t)}, (\mathbf{H}_{\text{Voxel}}^{(t)})_{v,j})}\left[\mathbb{E}_{\mathcal{N}}\left[\mathcal{L}_{\Phi}(\mathbf{Z}_{\mathbf{V}}^{(t)}, (\mathbf{H}_{\text{Voxel}}^{(t)})_{v,j}, \mathcal{N})\right]\right], \tag{7}$$

where $\mathcal{N}$ is the set of negative samples, $(\mathbf{Z}_{\mathbf{F}}^{(t)}, \boldsymbol{\psi}_{v,i}^{(t)})$ and $(\mathbf{Z}_{\mathbf{V}}^{(t)}, (\mathbf{H}_{\text{Voxel}}^{(t)})_{v,j})$ are the positive sample pairs, and $\mathcal{L}_{\Phi}$ is a standard $\texttt{Log-Softmax}$.

Table 1: Performance on brain classification: Mean ACC (%) $\pm$ standard deviation.

Methods	Bvfc	Bvfc-MEG	HCP-Mental	HCP-Age
Usad	$48.52_{\pm 1.94}$	$50.02_{\pm 1.13}$	$73.49_{\pm 1.56}$	$39.17_{\pm 1.68}$
HyperSAGCN	$51.92_{\pm 1.47}$	$51.19_{\pm 1.88}$	$90.37_{\pm 1.61}$	$47.38_{\pm 1.96}$
GMM	$53.11_{\pm 1.44}$	$53.04_{\pm 1.73}$	$90.92_{\pm 1.83}$	$47.75_{\pm 1.26}$
GraphMixer	$53.17_{\pm 1.21}$	$53.12_{\pm 1.18}$	$91.13_{\pm 1.44}$	$48.32_{\pm 1.11}$
BrainNetCnn	$49.10_{\pm 1.83}$	$50.12_{\pm 1.57}$	$83.58_{\pm 1.68}$	$42.26_{\pm 2.03}$
BrainGNN	$50.63_{\pm 1.67}$	$51.08_{\pm 0.96}$	$85.25_{\pm 2.17}$	$43.08_{\pm 1.54}$
FbNetGen	$50.18_{\pm 0.98}$	$50.94_{\pm 1.39}$	$84.47_{\pm 1.88}$	$42.83_{\pm 1.78}$
ADMire	$54.36_{\pm 1.39}$	$54.87_{\pm 1.92}$	$89.74_{\pm 1.93}$	$47.82_{\pm 1.72}$
PTGB	$55.89_{\pm 1.78}$	$55.11_{\pm 1.62}$	$92.58_{\pm 1.31}$	$48.41_{\pm 1.47}$
BNTransformer	$55.03_{\pm 1.35}$	$55.17_{\pm 1.74}$	$91.71_{\pm 1.48}$	$47.94_{\pm 1.15}$
BrainMixer	$\mathbf{67.24}_{\pm \mathbf{1.47}}$	$\mathbf{62.58}_{\pm \mathbf{1.12}}$	$\mathbf{96.32}_{\pm \mathbf{0.29}}$	$\mathbf{57.83}_{\pm \mathbf{1.03}}$

Data Augmentation & Negative Samples. MLP-Mixer-based architectures are known to have the potenial of overfitting (Liu et al., 2021). To mitigate this, we perform data augmentation. For $\mathcal{G}_F^{(t)} = (\mathcal{V}, \mathcal{E}^{(t)})$, in patch extraction, we randomly mask p connections and then we sample temporal walks to generate new patches. Note that, at the end, each patch is an induced subgraph and might include masked connections as well. Furthermore, to generate negative samples: ① To corrupt the functional connectivity, we randomly change one endpoint of a subset of connections. ② To corrupt the timeseries, we follow existing studies (Yue et al., 2022; Woo et al., 2022) on timeseries and replace a brain signal in time window t with another signal that is randomly selected from the batch.

Given a pre-trained model $\mathcal{M}$, for different downstream tasks in a semi-supervised setting, we fine-tune $\mathcal{M}$ using a small subset of labeled training data. Also, for each voxel, we concatenate its encodings from VA and FC Encoders.

4 Experiments

Dataset. We use six real-world datasets: ① We present Bvfc, a task-based fMRI large-scale dataset that includes voxel activity timeseries and functional connectivity of 3 subjects when looking at the 8460 images from 720 categories. This data is based on THINGS dataset (Hebart et al., 2023). ②Bvfc-MEG is the MEG counterpart of Bvfc. ③ ADHD (Milham et al., 2011) contains data for 250 subjects in the ADHD group and 450 subjects in the typically developed (TD) control group. ④ The Seizure detection TUH-EEG dataset (Shah et al., 2018) consists of EEG data (31 channels) of 642 subjects. ⑤ ASD (Craddock et al., 2013) contains data for 45 subjects in the ASD group and 45 subjects in the TD control group. ⑥ HCP (Van Essen et al., 2013) contains data from 7440 neuroimaging samples each of which is associated with one of the seven ground-truth mental states.

Evaluation Tasks. In our experiments we focus on 4 downstream tasks: ① Edge-Anomaly Detection (AD), ② Voxel AD, ③ Brain AD, and ④ Brain Classification. For the AD tasks, we follow previous studies (Behrouz & Seltzer, 2023a; Ma et al., 2021), and inject 1% and 5% anomalous edges into the functional connectivity in the control group of all datasets, except Bvfc, and Bvfc-MEG. Bvfc and Bvfc-MEG has ground-truth anomalies, the brain response of subjects when looking at not recognizable images, generated by generative adversarial neural network BigGAN (Brock et al., 2019). For brain classification, we focus on disease/disorder detection (in ADHD, ASD, and TUH-EEG), the category of seen object by the subject (in Bvfc, and Bvfc-MEG), and age prediction and mental state decoding (in HCP-Age, and HCP-Mental).

Baselines. For anomaly detection and graph classification tasks, we compare BrainMixer with state-of-the-art time series, graph, and brain anomaly detection and learning models: ① Graph-based methods: GOutlier (Aggarwal et al., 2011), NetWalk (Yu et al., 2018), HyperSAGCN (Zhang et al., 2020), Graph MLP-Mixer (GMM) (He et al., 2023), GraphMixer (Cong et al., 2023). ② brain-network-based methods: BrainGnn (Li et al., 2021), FbNetGen (Kan et al., 2022a), BrainNetCnn (Kawahara et al., 2017), ADMire (Behrouz & Seltzer, 2023b), and BNTransformer (Kan et al., 2022b), PTGB (Yang et al., 2023b). ③ Time-series-based methods: Usad (Audibert et al., 2020), Time Series Transformer (TST) (Zerveas et al., 2021), and Mvts (Potter et al., 2022). We may exclude some baselines in some tasks as they cannot be applied in that setting. The details of baselines can be found in Appendix F.1.

Table 2: Performance on anomaly detection: Mean AUC (%) ± standard deviation.

	Methods	BVFC	BVFC-MEG	HCP		ADHD		TUH-EEG		ASD	
	Anomaly %			1%	5 %	1%	5 %	1%	5 %	1%	5 %
Edge-level AD	GOUTLIER	65.12$_{\pm2.97}$	59.45$_{\pm2.61}$	62.47$_{\pm1.15}$	61.83$_{\pm1.28}$	65.37$_{\pm0.93}$	64.70$_{\pm2.09}$	65.61$_{\pm1.82}$	64.12$_{\pm0.97}$	60.85$_{\pm0.97}$	59.13$_{\pm1.86}$
	NETWALK	71.67$_{\pm1.56}$	62.75$_{\pm1.16}$	73.12$_{\pm1.25}$	72.19$_{\pm1.31}$	70.29$_{\pm2.15}$	69.86$_{\pm2.58}$	71.14$_{\pm1.36}$	70.27$_{\pm1.42}$	69.07$_{\pm2.20}$	68.52$_{\pm2.55}$
	HYPERSAGCN	80.17$_{\pm1.59}$	70.83$_{\pm1.27}$	82.94$_{\pm1.14}$	81.98$_{\pm1.58}$	84.22$_{\pm1.61}$	83.96$_{\pm1.47}$	73.99$_{\pm0.83}$	72.65$_{\pm0.97}$	73.26$_{\pm1.08}$	73.18$_{\pm0.92}$
	GRAPHMIXER	87.13$_{\pm0.99}$	75.91$_{\pm1.59}$	86.87$_{\pm1.96}$	86.19$_{\pm1.48}$	85.12$_{\pm1.46}$	84.86$_{\pm1.58}$	75.93$_{\pm0.95}$	75.12$_{\pm1.08}$	84.91$_{\pm2.27}$	83.52$_{\pm2.03}$
	BRAINNETCNN	80.92$_{\pm1.18}$	71.54$_{\pm2.07}$	80.79$_{\pm1.23}$	79.44$_{\pm1.18}$	80.58$_{\pm1.62}$	79.95$_{\pm2.01}$	73.06$_{\pm1.74}$	72.87$_{\pm1.31}$	72.68$_{\pm2.12}$	72.01$_{\pm1.45}$
	BRAINGNN	81.96$_{\pm1.76}$	72.68$_{\pm1.13}$	82.15$_{\pm1.84}$	81.38$_{\pm1.61}$	79.02$_{\pm1.85}$	78.64$_{\pm1.43}$	72.96$_{\pm1.58}$	71.73$_{\pm1.14}$	72.14$_{\pm1.25}$	71.82$_{\pm1.73}$
	FBNETGEN	81.58$_{\pm1.92}$	72.66$_{\pm1.52}$	82.05$_{\pm1.19}$	81.53$_{\pm1.82}$	79.89$_{\pm1.63}$	78.97$_{\pm1.84}$	73.04$_{\pm1.53}$	72.56$_{\pm1.33}$	72.51$_{\pm1.28}$	71.62$_{\pm1.82}$
	ADMIRE	87.12$_{\pm1.61}$	75.91$_{\pm1.43}$	87.01$_{\pm1.27}$	86.38$_{\pm1.17}$	86.23$_{\pm1.74}$	85.18$_{\pm2.21}$	76.68$_{\pm1.82}$	75.14$_{\pm1.67}$	86.52$_{\pm1.72}$	85.44$_{\pm1.49}$
	PTGB	86.52$_{\pm1.64}$	75.93$_{\pm1.71}$	86.83$_{\pm1.59}$	86.00$_{\pm1.28}$	86.14$_{\pm1.15}$	85.22$_{\pm1.21}$	75.98$_{\pm1.16}$	74.92$_{\pm1.08}$	86.18$_{\pm1.58}$	85.72$_{\pm1.05}$
	BNTRANSFORMER	86.61$_{\pm1.72}$	75.82$_{\pm1.18}$	86.22$_{\pm1.77}$	85.15$_{\pm1.12}$	85.83$_{\pm1.97}$	85.14$_{\pm1.67}$	75.91$_{\pm1.72}$	75.24$_{\pm1.53}$	74.92$_{\pm1.18}$	74.11$_{\pm1.37}$
	BRAINMIXER	**91.62$_{\pm1.36}$**	**82.58$_{\pm1.92}$**	**90.14$_{\pm1.72}$**	**90.02$_{\pm1.49}$**	**91.74$_{\pm0.93}$**	**91.48$_{\pm1.41}$**	**80.91$_{\pm1.19}$**	**80.85$_{\pm1.62}$**	**90.44$_{\pm1.57}$**	**90.27$_{\pm1.39}$**
Voxel-level AD	USAD	68.27$_{\pm1.16}$	62.73$_{\pm1.27}$	65.49$_{\pm1.31}$	65.01$_{\pm1.18}$	72.79$_{\pm1.48}$	72.19$_{\pm0.94}$	72.81$_{\pm1.42}$	71.36$_{\pm1.03}$	66.28$_{\pm1.16}$	65.17$_{\pm1.15}$
	TST	70.62$_{\pm1.48}$	68.57$_{\pm1.81}$	69.18$_{\pm1.64}$	69.11$_{\pm1.32}$	74.81$_{\pm1.14}$	73.99$_{\pm1.47}$	73.71$_{\pm1.55}$	73.03$_{\pm1.47}$	69.23$_{\pm1.82}$	68.94$_{\pm1.73}$
	MVTS	N/A	N/A	N/A	N/A	N/A	N/A	80.99$_{\pm1.36}$	80.27$_{\pm1.49}$	N/A	N/A
	GOUTLIER	64.66$_{\pm2.38}$	60.17$_{\pm1.25}$	63.59$_{\pm1.62}$	63.07$_{\pm1.52}$	68.97$_{\pm1.16}$	67.12$_{\pm0.93}$	65.18$_{\pm1.09}$	65.01$_{\pm1.57}$	59.67$_{\pm1.42}$	58.49$_{\pm1.35}$
	NETWALK	68.73$_{\pm1.16}$	63.61$_{\pm1.31}$	66.98$_{\pm1.44}$	66.04$_{\pm1.63}$	75.16$_{\pm1.23}$	74.73$_{\pm1.01}$	72.21$_{\pm0.91}$	71.62$_{\pm1.46}$	71.28$_{\pm1.17}$	71.02$_{\pm1.49}$
	HYPERSAGCN	78.84$_{\pm1.22}$	71.62$_{\pm1.96}$	80.74$_{\pm1.51}$	79.18$_{\pm1.83}$	83.94$_{\pm1.13}$	83.01$_{\pm0.92}$	75.62$_{\pm1.12}$	74.83$_{\pm0.78}$	74.93$_{\pm1.47}$	74.15$_{\pm1.19}$
	GRAPHMIXER	76.94$_{\pm1.68}$	71.44$_{\pm1.39}$	81.55$_{\pm1.82}$	81.07$_{\pm1.27}$	81.37$_{\pm1.09}$	80.83$_{\pm1.16}$	72.95$_{\pm1.26}$	72.01$_{\pm0.82}$	72.49$_{\pm1.28}$	72.27$_{\pm1.69}$
	BRAINNETCNN	80.17$_{\pm1.49}$	73.91$_{\pm1.54}$	82.75$_{\pm1.27}$	82.21$_{\pm1.73}$	82.79$_{\pm1.08}$	81.12$_{\pm1.16}$	73.98$_{\pm1.24}$	73.01$_{\pm1.08}$	73.18$_{\pm0.95}$	72.88$_{\pm1.04}$
	BRAINGNN	79.92$_{\pm1.63}$	73.25$_{\pm1.94}$	82.99$_{\pm1.65}$	82.13$_{\pm1.66}$	81.14$_{\pm1.05}$	80.83$_{\pm0.87}$	73.06$_{\pm1.14}$	72.74$_{\pm0.86}$	72.54$_{\pm1.38}$	71.12$_{\pm1.19}$
	FBNETGEN	79.17$_{\pm2.04}$	72.35$_{\pm1.84}$	82.26$_{\pm1.37}$	81.62$_{\pm1.49}$	80.91$_{\pm1.12}$	80.94$_{\pm1.74}$	72.53$_{\pm1.48}$	72.06$_{\pm1.29}$	72.11$_{\pm1.94}$	71.28$_{\pm1.22}$
	PTGB	85.18$_{\pm1.83}$	76.16$_{\pm1.08}$	85.72$_{\pm1.14}$	84.95$_{\pm1.33}$	86.43$_{\pm1.16}$	86.36$_{\pm1.15}$	77.54$_{\pm1.37}$	77.32$_{\pm1.21}$	77.92$_{\pm1.26}$	77.76$_{\pm1.25}$
	BN-TRANSFORMER	85.19$_{\pm1.23}$	75.67$_{\pm1.14}$	85.02$_{\pm0.96}$	84.36$_{\pm1.59}$	86.13$_{\pm1.21}$	86.11$_{\pm1.82}$	77.96$_{\pm1.32}$	77.08$_{\pm1.06}$	76.05$_{\pm1.52}$	75.72$_{\pm1.18}$
	BRAINMIXER	**90.14$_{\pm1.57}$**	**81.52$_{\pm1.32}$**	**89.27$_{\pm1.61}$**	**88.94$_{\pm1.24}$**	**89.97$_{\pm1.14}$**	**89.81$_{\pm1.27}$**	**79.45$_{\pm1.19}$**	**79.23$_{\pm0.94}$**	**89.51$_{\pm1.78}$**	**89.24$_{\pm1.59}$**
Brain-level AD	USAD	71.93$_{\pm1.15}$	61.32$_{\pm1.71}$	67.79$_{\pm2.28}$	67.36$_{\pm2.61}$	82.87$_{\pm2.03}$	80.52$_{\pm1.84}$	72.03$_{\pm1.17}$	71.48$_{\pm1.05}$	71.62$_{\pm1.58}$	70.98$_{\pm1.41}$
	TST	72.47$_{\pm1.23}$	67.12$_{\pm2.07}$	67.94$_{\pm1.69}$	67.22$_{\pm1.17}$	83.54$_{\pm1.38}$	83.04$_{\pm1.12}$	72.96$_{\pm1.39}$	72.11$_{\pm1.58}$	72.76$_{\pm1.71}$	72.04$_{\pm1.56}$
	MVTS	N/A	N/A	N/A	N/A	N/A	N/A	83.53$_{\pm1.91}$	82.41$_{\pm1.02}$	N/A	N/A
	NETWALK	72.16$_{\pm1.44}$	69.57$_{\pm1.73}$	69.14$_{\pm1.49}$	68.66$_{\pm1.52}$	83.11$_{\pm1.02}$	82.81$_{\pm1.61}$	71.06$_{\pm1.05}$	69.94$_{\pm1.12}$	72.85$_{\pm1.17}$	72.21$_{\pm1.34}$
	HYPERSAGCN	80.25$_{\pm1.15}$	76.91$_{\pm1.18}$	72.26$_{\pm1.47}$	72.01$_{\pm1.21}$	86.94$_{\pm1.63}$	86.17$_{\pm1.49}$	75.31$_{\pm0.85}$	74.79$_{\pm1.09}$	76.72$_{\pm1.32}$	75.81$_{\pm1.58}$
	GMM	81.79$_{\pm1.24}$	77.84$_{\pm1.52}$	74.87$_{\pm1.58}$	74.02$_{\pm1.10}$	85.89$_{\pm0.98}$	85.03$_{\pm1.18}$	76.62$_{\pm1.17}$	76.11$_{\pm1.26}$	76.37$_{\pm1.83}$	75.68$_{\pm1.59}$
	GRAPHMIXER	82.56$_{\pm1.19}$	77.91$_{\pm1.26}$	75.03$_{\pm1.72}$	74.46$_{\pm1.53}$	86.02$_{\pm1.15}$	85.64$_{\pm1.09}$	77.49$_{\pm1.09}$	76.63$_{\pm1.22}$	76.82$_{\pm1.84}$	76.18$_{\pm1.80}$
	BRAINNETCNN	78.47$_{\pm1.18}$	73.12$_{\pm1.27}$	70.73$_{\pm1.77}$	70.12$_{\pm1.86}$	85.84$_{\pm0.96}$	85.07$_{\pm1.52}$	73.92$_{\pm0.97}$	73.07$_{\pm1.51}$	75.96$_{\pm1.66}$	75.03$_{\pm1.28}$
	BRAINGNN	79.81$_{\pm1.57}$	75.28$_{\pm1.61}$	72.98$_{\pm1.55}$	72.41$_{\pm1.16}$	84.59$_{\pm1.26}$	83.72$_{\pm1.35}$	72.41$_{\pm1.38}$	71.55$_{\pm1.16}$	75.12$_{\pm1.33}$	74.57$_{\pm1.52}$
	FBNETGEN	78.94$_{\pm1.24}$	74.49$_{\pm1.33}$	71.62$_{\pm1.53}$	71.06$_{\pm1.48}$	84.67$_{\pm1.26}$	84.08$_{\pm1.37}$	72.69$_{\pm1.18}$	71.87$_{\pm1.12}$	75.34$_{\pm1.21}$	74.73$_{\pm1.39}$
	ADMIRE	83.72$_{\pm1.18}$	78.83$_{\pm1.56}$	75.52$_{\pm1.81}$	74.59$_{\pm1.12}$	86.27$_{\pm1.72}$	85.18$_{\pm1.56}$	78.12$_{\pm1.47}$	77.59$_{\pm1.68}$	77.18$_{\pm1.61}$	76.33$_{\pm1.45}$
	PTGB	84.08$_{\pm1.35}$	79.68$_{\pm1.62}$	76.01$_{\pm1.07}$	75.13$_{\pm1.48}$	87.59$_{\pm1.12}$	86.99$_{\pm0.96}$	79.17$_{\pm1.36}$	78.64$_{\pm1.55}$	80.56$_{\pm1.29}$	80.04$_{\pm1.16}$
	BN-TRANSFORMER	83.86$_{\pm1.52}$	79.03$_{\pm1.78}$	75.64$_{\pm1.82}$	75.09$_{\pm1.18}$	87.54$_{\pm1.04}$	86.92$_{\pm1.48}$	79.36$_{\pm1.71}$	78.08$_{\pm1.16}$	77.19$_{\pm2.01}$	76.58$_{\pm1.73}$
	BRAINMIXER	**88.13$_{\pm1.27}$**	**84.59$_{\pm1.70}$**	**80.67$_{\pm1.13}$**	**80.49$_{\pm1.07}$**	**91.38$_{\pm0.94}$**	**90.98$_{\pm1.02}$**	**85.74$_{\pm1.16}$**	**85.63$_{\pm1.23}$**	**89.14$_{\pm1.54}$**	**88.99$_{\pm1.15}$**

Table 3: Ablation study on BRAINMIXER. AUC scores on edge AD and ACC on classification.

Methods	BVFC		BVFC-MEG		HCP		ADHD	
	Edge AD	Classification	Edge AD	Classification	Edge AD	Classification	Edge AD	Classification
BRAINMIXER	**91.62$_{\pm1.36}$**	**67.24**$_{\pm1.47}$	**82.58$_{\pm1.92}$**	**62.68$_{\pm1.12}$**	**90.02$_{\pm1.49}$**	**96.32$_{\pm0.29}$**	**91.48$_{\pm1.41}$**	**90.98$_{\pm1.02}$**
Without Pre-training	88.75$_{\pm2.16}$	63.58$_{\pm2.09}$	80.21$_{\pm1.63}$	61.02$_{\pm1.37}$	88.14$_{\pm1.29}$	93.81$_{\pm0.92}$	90.18$_{\pm1.13}$	89.27$_{\pm1.06}$
Without VA Encoder	87.99$_{\pm2.04}$	59.14$_{\pm4.51}$	78.52$_{\pm2.18}$	60.53$_{\pm1.83}$	86.97$_{\pm2.05}$	92.41$_{\pm1.24}$	88.29$_{\pm1.41}$	88.76$_{\pm1.19}$
Without FC Encoder	84.27$_{\pm4.37}$	65.82$_{\pm2.18}$	77.09$_{\pm3.41}$	59.73$_{\pm1.12}$	85.59$_{\pm2.47}$	91.64$_{\pm1.58}$	86.97$_{\pm1.16}$	87.62$_{\pm2.16}$
Without Functional Patching	86.35$_{\pm2.97}$	60.42$_{\pm3.53}$	77.21$_{\pm1.93}$	60.28$_{\pm1.72}$	86.14$_{\pm3.09}$	91.97$_{\pm1.88}$	87.51$_{\pm1.86}$	88.25$_{\pm2.53}$
Replace TPMIXER by MEAN(.)	88.51$_{\pm1.03}$	63.38$_{\pm1.48}$	78.94$_{\pm1.85}$	60.91$_{\pm2.01}$	87.52$_{\pm1.91}$	93.31$_{\pm1.73}$	89.04$_{\pm0.95}$	89.11$_{\pm1.52}$
Static Self-Attention	88.39$_{\pm1.40}$	63.01$_{\pm2.10}$	78.63$_{\pm1.97}$	60.78$_{\pm1.64}$	87.04$_{\pm1.53}$	92.95$_{\pm1.49}$	88.96$_{\pm1.22}$	88.83$_{\pm2.07}$
Remove Time Encoding	89.58$_{\pm0.81}$	66.14$_{\pm1.52}$	79.91$_{\pm1.75}$	61.19$_{\pm1.36}$	88.82$_{\pm2.07}$	94.12$_{\pm1.92}$	90.57$_{\pm0.91}$	89.99$_{\pm1.04}$
fix $\theta = 0$	83.60$_{\pm4.52}$	59.33$_{\pm2.58}$	75.96$_{\pm2.05}$	59.11$_{\pm1.46}$	85.39$_{\pm1.52}$	90.51$_{\pm1.38}$	86.24$_{\pm2.01}$	87.18$_{\pm1.94}$

Brain Classification. Table 1 reports the performance of BRAINMIXER and baselines on brain classification tasks. BRAINMIXER achieves the best accuracy on all datasets with 14.3% average improvement (20.3% best improvement) over the best baseline. There are three main reasons for BRAINMIXER's superior performance: ① While the time series-based model only uses voxel activity timeseries, and graph-based methods only use functional connectivity graph, BRAINMIXER takes advantage of both and learns the brain representation at different levels of granularity, which can provide complementary information. ② Static methods (e.g., PTGB, BRAINGNN, etc.), miss the dynamics of brain activity, while BRAINMIXER employs a time encoding module to learn temporal properties. ③ Compared to graph learning methods (e.g., GMM, GRAPHMIXER, etc.), BRAINMIXER is specifically designed for the brain, taking advantage of its special properties.

Anomaly Detection. Table 2 reports the performance of BRAINMIXER and baselines on anomaly detection tasks at different scales: i.e., edge-, voxel-, and brain-level. BRAINMIXER achieves the best AUC on all datasets with 6.2%, 5.7%, 4.81% average improvement over the best baseline in edge AD, voxel AD, and brain AD, respectively. The main reasons for this superior performance are as above. Note that brain-level anomaly detection can also be seen as a brain classification task. However, here, based on the nature of the data, we separate these two tasks.

Ablation Study. We next conduct ablation studies on the BVFC, BVFC-MEG, HCP, and ADHD datasets to validate the effectiveness of BRAINMIXER's critical components. Table 3 shows AUC for edge AD and accuracy for classification tasks. The first row reports the performance of the complete BRAINMIXER implementation with pre-training. Each subsequent row shows results for

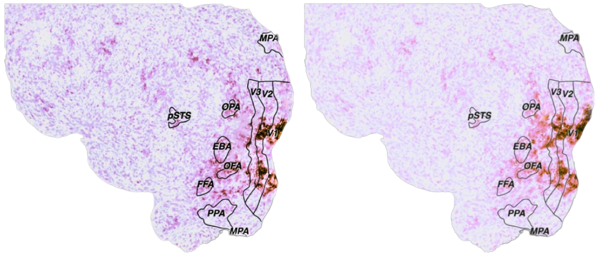

Figure 2: Average distribution of brain activities in the visual cortex when seeing (Left) GAN-generated images, (Right) Normal image.

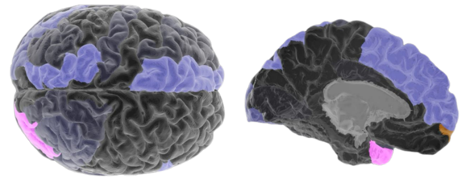

Figure 3: The distribution of detected abnormal voxels by BrainMixer in condition ADHD group.

BrainMixer with one module modification: row 2 removes the pre-training phase, row 3 removes the VA Encoder module, row 4 removes FC Encoder module, row 5 removes functional patching and randomly patches voxels, row replaces TPMixer with Mean(.) pooling, row 7 replaces dynamic with static self-attention, row 8 removes time encoder, the last row set $\theta = 0$, removing biased in the sampling. These results show that each component is critical for achieving BrainMixer's superior performance. The greatest contribution comes from biased sampling, VA and FC encoders, functional patching, and dynamic self-attention, respectively.

Parameter Sensitivity. We discuss the effect of the number of walks, M, the walk length, m, and time decay, θ on the performance in Appendix G. Results show that increasing the number of walks results in better performance as each patch is a better representation of the node's neighborhood. The effect of the walk length on performance peaks at a certain point, but the exact value varies with datasets. In Appendix G, we further discuss how aggregating timeseries to obtain beta weights and aggregating voxels to obtain ROIs can affect performance.

How Does Brain Detect GAN Generated Images? The visual cortex, responsible for processing visual information, is hierarchically organized with multiple layers building upon simpler features at lower stages (Van Essen & Maunsell, 1983). Initially, neurons detect edges and colors, but on deeper levels, they specialize in recognizing more complex patterns and objects. Figure 2 (Left) (resp. (Right)) reports the average distribution of brain activity of a subject when looking at non-recognizable images (resp. natural images). Interestingly, while the distributions share similar patterns in lower levels (e.g., V1 and V2 voxels), higher-level voxels (e.g., V3) are less active when the subject sees non-recognizable images.

Case Study: ADHD In this case study, we train our model on the neuroimages of the typically developed group and test it on the ADHD condition group to detect abnormal voxel activities that might be correlated to ADHD symptoms. Figure 3 reports the distribution of anomalous voxels within the brain of the ADHD group. 78% of all found abnormal voxel activities by BrainMixer are located in the Frontal Pole, Left and Right Temporal Poles, and Lingual Gyrus. Surprisingly, these findings are consistent with previous studies on ADHD, which use diffusion tensor imaging (Lei et al., 2014) and Forman–Ricci curvature changes (Chatterjee et al., 2021).

5 Conclusion

In this work, we present an unsupervised pre-training framework, BrainMixer, that bridges the representation learning of voxel activity and functional connectivity by maximizing their mutual information. BrainMixer presents two novel variations of MLP-Mixer to multivariate timeseries (VA Encoder) and graphs (FC Encoder) that both take advantage of special properties of the brain to obtain effective representations of voxels. Consequently, the experimental results show the potential of BrainMixer in ① detecting abnormal brain activity that might cause a brain disease/disorder, ② disease/disorder detection, and ③ understanding object representation in the brain. Experiments further support the significance of each BrainMixer's component and show its superior performance compared to the state-of-the-art in a variety of tasks. We discuss potential limitations and future work in Appendix H.

References

Charu C. Aggarwal, Yuchen Zhao, and Philip S. Yu. Outlier detection in graph streams. In *2011 IEEE 27th International Conference on Data Engineering*, pp. 399–409, 2011. doi: 10.1109/ICDE.2011.5767885.

Julien Audibert, Pietro Michiardi, Frédéric Guyard, Sébastien Marti, and Maria A Zuluaga. Usad: Unsupervised anomaly detection on multivariate time series. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 3395–3404, 2020.

Lev Kiar Avberšek and Grega Repovš. Deep learning in neuroimaging data analysis: applications, challenges, and solutions. *Frontiers in neuroimaging*, 1:981642, 2022.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016.

Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/ddf354219aac374f1d40b7e760ee5bb7-Paper.pdf.

Ali Behrouz and Margo Seltzer. Anomaly detection in multiplex dynamic networks: from blockchain security to brain disease prediction. In *NeurIPS 2022 Temporal Graph Learning Workshop*, 2022. URL https://openreview.net/forum?id=UDGZDfwmay.

Ali Behrouz and Margo Seltzer. Anomaly detection in human brain via inductive learning on temporal multiplex networks. In *Machine Learning for Healthcare Conference*, volume 219. PMLR, 2023a.

Ali Behrouz and Margo Seltzer. ADMIRE++: Explainable anomaly detection in the human brain via inductive learning on temporal multiplex networks. In *ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH)*, 2023b. URL https://openreview.net/forum?id=t4H8acYudJ.

Ali Behrouz, Farnoosh Hashemi, Sadaf Sadeghian, and Margo Seltzer. CAT-walk: Inductive hypergraph learning via set walks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=QG4nJBNEar.

Signe L Bray, Catie Chang, and Fumiko Hoeft. Applications of multivariate pattern classification analyses in developmental neuroimaging of healthy and clinical populations. *Frontiers in human neuroscience*, 3:898, 2009.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1xsqj09Fm.

Donghong Cai, Junru Chen, Yang Yang, Teng Liu, and Yafeng Li. Mbrain: A multi-channel self-supervised learning framework for brain signals. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, pp. 130–141, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599426. URL https://doi.org/10.1145/3580305.3599426.

Vince D. Calhoun, Robyn Miller, Godfrey Pearlson, and Tulay Adalı. The chronnectome: Time-varying connectivity networks as the next frontier in fmri data discovery. *Neuron*, 84(2):262–274, 2014. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2014.10.015. URL https://www.sciencedirect.com/science/article/pii/S0896627314009131.

Martin J Chadwick, Heidi M Bonnici, and Eleanor A Maguire. Decoding information in the human hippocampus: a user's guide. *Neuropsychologia*, 50(13):3107–3121, 2012.

Benjamin Paul Chamberlain, Sergey Shirobokov, Emanuele Rossi, Fabrizio Frasca, Thomas Markovich, Nils Yannick Hammerla, Michael M. Bronstein, and Max Hansmire. Graph neural networks for link prediction with subgraph sketching. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=m1oqEOAozQU.

Tanima Chatterjee, Réka Albert, Stuti Thapliyal, Nazanin Azarhooshang, and Bhaskar DasGupta. Detecting network anomalies using forman–ricci curvature and a case study for human brain networks. *Scientific Reports*, 11(1):8121, Apr 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-87587-z. URL https://doi.org/10.1038/s41598-021-87587-z.

Gang Chen, B Douglas Ward, Chunming Xie, Wenjun Li, Zhilin Wu, Jennifer L Jones, Malgorzata Franczak, Piero Antuono, and Shi-Jiang Li. Classification of alzheimer disease, mild cognitive impairment, and normal cognitive status with large-scale network analysis based on resting-state functional mr imaging. *Radiology*, 259(1):213, 2011.

Si-An Chen, Chun-Liang Li, Nate Yoder, Sercan O Arik, and Tomas Pfister. Tsmixer: An all-mlp architecture for time series forecasting. *arXiv preprint arXiv:2303.06053*, 2023.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

Weilin Cong, Si Zhang, Jian Kang, Baichuan Yuan, Hao Wu, Xin Zhou, Hanghang Tong, and Mehrdad Mahdavi. Do we really need complicated model architectures for temporal networks? In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ayPPc0SyLv1.

Aurelio Cortese, Saori C Tanaka, Kaoru Amano, Ai Koizumi, Hakwan Lau, Yuka Sasaki, Kazuhisa Shibata, Vincent Taschereau-Dumouchel, Takeo Watanabe, and Mitsuo Kawato. The decnef collection, fmri data from closed-loop decoded neurofeedback experiments. *Scientific data*, 8(1): 65, 2021.

Cameron Craddock, Yassine Benhajali, Carlton Chu, Francois Chouinard, Alan Evans, András Jakab, Budhachandra Singh Khundrakpam, John David Lewis, Qingyang Li, Michael Milham, et al. The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. *Frontiers in Neuroinformatics*, 7:27, 2013.

Alexander Craik, Yongtian He, and Jose L Contreras-Vidal. Deep learning for electroencephalogram (eeg) classification tasks: a review. *Journal of neural engineering*, 16(3):031001, 2019.

Hejie Cui, Wei Dai, Yanqiao Zhu, Xiaoxiao Li, Lifang He, and Carl Yang. Interpretable graph neural networks for connectome-based brain disorder analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 375–385. Springer, 2022.

Fernando Lopes da Silva. Neural mechanisms underlying brain waves: from neural membranes to networks. *Electroencephalography and clinical neurophysiology*, 79(2):81–93, 1991.

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. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Yuhui Du, Zening Fu, and Vince D Calhoun. Classification and prediction of brain disorders using functional connectivity: promising but challenging. *Frontiers in neuroscience*, 12:525, 2018.

Javier Gonzalez-Castillo and Peter A Bandettini. Task-based dynamic functional connectivity: Recent findings and open questions. *Neuroimage*, 180:526–533, 2018.

Michael Greicius. Resting-state functional connectivity in neuropsychiatric disorders. *Current opinion in neurology*, 21(4):424–430, 2008.

Yanming Guo, Yu Liu, Ard Oerlemans, Songyang Lao, Song Wu, and Michael S Lew. Deep learning for visual understanding: A review. *Neurocomputing*, 187:27–48, 2016.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 297–304. JMLR Workshop and Conference Proceedings, 2010.

Xiaoxin He, Bryan Hooi, Thomas Laurent, Adam Perold, Yann LeCun, and Xavier Bresson. A generalization of vit/mlp-mixer to graphs. In *International Conference on Machine Learning*, pp. 12724–12745. PMLR, 2023.

Martin N Hebart, Oliver Contier, Lina Teichmann, Adam H Rockter, Charles Y Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I Baker. Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *Elife*, 12:e82580, 2023.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus), 2020.

Leanna M Hernandez, Jeffrey D Rudie, Shulamite A Green, Susan Bookheimer, and Mirella Dapretto. Neural signatures of autism spectrum disorders: insights into brain network dynamics. *Neuropsychopharmacology*, 40(1):171–189, 2015.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bklr3j0cKX.

Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature communications*, 8(1):15037, 2017.

Yang Hu, Haoxuan You, Zhecan Wang, Zhicheng Wang, Erjin Zhou, and Yue Gao. Graph-mlp: Node classification without message passing in graph. *arXiv preprint arXiv:2106.04051*, 2021.

R Matthew Hutchison, Thilo Womelsdorf, Elena A Allen, Peter A Bandettini, Vince D Calhoun, Maurizio Corbetta, Stefania Della Penna, Jeff H Duyn, Gary H Glover, Javier Gonzalez-Castillo, et al. Dynamic functional connectivity: promise, issues, and interpretations. *Neuroimage*, 80: 360–378, 2013.

Yuli Jiang, Yu Rong, Hong Cheng, Xin Huang, Kangfei Zhao, and Junzhou Huang. Query driven-graph neural networks for community search: From non-attributed, attributed, to interactive attributed, 2021. URL https://arxiv.org/abs/2104.03583.

Biao Jie, Mingxia Liu, Xi Jiang, and Daoqiang Zhang. Sub-network based kernels for brain network classification. In *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 622–629, 2016.

Xuan Kan, Hejie Cui, Joshua Lukemire, Ying Guo, and Carl Yang. Fbnetgen: Task-aware gnn-based fmri analysis via functional brain network generation. In *International Conference on Medical Imaging with Deep Learning*, pp. 618–637. PMLR, 2022a.

Xuan Kan, Wei Dai, Hejie Cui, Zilong Zhang, Ying Guo, and Carl Yang. Brain network transformer. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022b. URL https://openreview.net/forum?id=1cJ1cbA6NLN.

Jeremy Kawahara, Colin J Brown, Steven P Miller, Brian G Booth, Vann Chau, Ruth E Grunau, Jill G Zwicker, and Ghassan Hamarneh. Brainnetcnn: Convolutional neural networks for brain networks; towards predicting neurodevelopment. *NeuroImage*, 146:1038–1049, 2017.

Tommaso Lanciano, Francesco Bonchi, and Aristides Gionis. Explainable classification of brain networks via contrast subgraphs. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery; Data Mining*, KDD '20, pp. 3308–3318, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3403383. URL https://doi.org/10.1145/3394486.3403383.

Sue-Hyun Lee and Chris I Baker. Multi-voxel decoding and the topography of maintained information during visual working memory. *Frontiers in systems neuroscience*, 10:2, 2016.

Du Lei, Jun Ma, Xiaoxia Du, Guohua Shen, Xingming Jin, and Qiyong Gong. Microstructural abnormalities in the combined and inattentive subtypes of attention deficit hyperactivity disorder: a diffusion tensor imaging study. *Scientific reports*, 4(1):6875, 2014.

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. *Advances in neural information processing systems*, 32, 2019.

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*, 74:102233, 2021.

Zhe Li, Zhongwen Rao, Lujia Pan, and Zenglin Xu. Mts-mixers: Multivariate time series forecasting via factorized temporal and channel mixing. *arXiv preprint arXiv:2302.04501*, 2023.

Hanxiao Liu, Zihang Dai, David So, and Quoc V Le. Pay attention to mlps. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 9204–9215. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/4cc05b35c2f937c5bd9e7d41d3686fff-Paper.pdf.

Xiaoxiao Ma, Jia Wu, Shan Xue, Jian Yang, Chuan Zhou, Quan Z. Sheng, Hui Xiong, and Leman Akoglu. A comprehensive survey on graph anomaly detection with deep learning. *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2021. doi: 10.1109/TKDE.2021.3118815.

Abdelhak Mahmoudi, Sylvain Takerkart, Fakhita Regragui, Driss Boussaoud, Andrea Brovelli, et al. Multivoxel pattern analysis for fmri data: a review. *Computational and mathematical methods in medicine*, 2012, 2012.

Chris McNorgan, Gregory J Smith, and Erica S Edwards. Integrating functional connectivity and mvpa through a multiple constraint network analysis. *Neuroimage*, 208:116412, 2020.

Michael P. Milham, Jan Buitelaar, F. Xavier Castellanos, Daniel Dickstein, Damien Fair, David Kennedy, Beatric Luna, Michael P. Milham, Stewart Mostofsky, Joel Nigg, Julie B. Schweitzer, Katerina Velanova, Yu-Feng Wang, and Yu-Feng Zang. 1000 functional connectome project. *1000 Functional Connectome Project*, 1, July 2011.

Tom M Mitchell, Rebecca Hutchinson, Marcel Adam Just, Radu S Niculescu, Francisco Pereira, and Xuerui Wang. Classifying instantaneous cognitive states from fmri data. In *AMIA annual symposium proceedings*, volume 2003, pp. 465. American Medical Informatics Association, 2003.

Alfonso Nieto-Castanon. Brain-wide connectome inferences using functional connectivity multivariate pattern analyses (fc-mvpa). *PLoS Computational Biology*, 18(11):e1010634, 2022.

Youngser Park, C Priebe, D Marchette, and Abdou Youssef. Anomaly detection using scan statistics on time series hypergraphs. In *Link Analysis, Counterterrorism and Security (LACTS) Conference*, pp. 9. SIAM Pennsylvania, 2009.

Russell A. Poldrack and Martha J. Farah. Progress and challenges in probing the human brain. *Nature*, 526(7573):371–379, Oct 2015. ISSN 1476-4687. doi: 10.1038/nature15692. URL https://doi.org/10.1038/nature15692.

Russell A Poldrack and Krzysztof J Gorgolewski. Making big data open: data sharing in neuroimaging. *Nature neuroscience*, 17(11):1510–1517, 2014.

İlkay Yıldız Potter, George Zerveas, Carsten Eickhoff, and Dominique Duncan. Unsupervised multivariate time-series transformers for seizure identification on eeg. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 1304–1311. IEEE, 2022.

Raimon HR Pruim, Maarten Mennes, Daan van Rooij, Alberto Llera, Jan K Buitelaar, and Christian F Beckmann. Ica-aroma: A robust ica-based strategy for removing motion artifacts from fmri data. *Neuroimage*, 112:267–277, 2015.

Zvi N. Roth and Elisha P. Merriam. Representations in human primary visual cortex drift over time. *Nature Communications*, 14(1):4422, Jul 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-40144-w. URL https://doi.org/10.1038/s41467-023-40144-w.

Zvi N. Roth, Kendrick Kay, and Elisha P. Merriam. Natural scene sampling reveals reliable coarse-scale orientation tuning in human v1. *Nature Communications*, 13(1):6469, Oct 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-34134-7. URL https://doi.org/10.1038/s41467-022-34134-7.

Ramit Sawhney, Shivam Agarwal, Arnav Wadhwa, Tyler Derr, and Rajiv Ratn Shah. Stock selection via spatiotemporal hypergraph attention network: A learning to rank approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 497–504, 2021.

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. *Cereb Cortex*, 28(9):3095–3114, September 2018.

Steffen Schneider, Jin Hwa Lee, and Mackenzie Weygandt Mathis. Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, pp. 1–9, 2023.

Vinit Shah, Eva Von Weltin, Silvia Lopez, James Riley McHugh, Lillian Veloso, Meysam Golmohammadi, Iyad Obeid, and Joseph Picone. The temple university hospital seizure detection corpus. *Frontiers in neuroinformatics*, 12:83, 2018.

Afshin Shoeibi, Marjane Khodatars, Navid Ghassemi, Mahboobeh Jafari, Parisa Moridian, Roohallah Alizadehsani, Maryam Panahiazar, Fahime Khozeimeh, Assef Zare, Hossein Hosseini-Nejad, et al. Epileptic seizures detection using deep learning techniques: A review. *International Journal of Environmental Research and Public Health*, 18(11):5780, 2021.

Stephen M Smith, Diego Vidaurre, Christian F Beckmann, Matthew F Glasser, Mark Jenkinson, Karla L Miller, Thomas E Nichols, Emma C Robinson, Gholamreza Salimi-Khorshidi, Mark W Woolrich, Deanna M Barch, Kamil Uğurbil, and David C Van Essen. Functional connectomics from resting-state fMRI. *Trends Cogn Sci*, 17(12):666–682, November 2013.

B Sundermann, D Herr, W Schwindt, and B Pfleiderer. Multivariate classification of blood oxygen level–dependent fmri data with diagnostic intention: a clinical perspective. *American Journal of neuroradiology*, 35(5):848–855, 2014.

Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Peter Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. MLP-mixer: An all-MLP architecture for vision. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=EI2KOXKdnP.

Lucina Q Uddin, DR Dajani, W Voorhies, H Bednarz, and RK Kana. Progress and roadblocks in the search for brain-based biomarkers of autism and attention-deficit/hyperactivity disorder. *Translational psychiatry*, 7(8):e1218–e1218, 2017.

Martijn P Van Den Heuvel and Hilleke E Hulshoff Pol. Exploring the brain network: a review on resting-state fmri functional connectivity. *European neuropsychopharmacology*, 20(8):519–534, 2010.

David C Van Essen and John HR Maunsell. Hierarchical organization and functional streams in the visual cortex. *Trends in neurosciences*, 6:370–375, 1983.

David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al. The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.

Eliana Vassena, James Deraeve, and William H. Alexander. Surprise, value and control in anterior cingulate cortex during speeded decision-making. *Nature Human Behaviour*, 4(4):412–422, Apr 2020. ISSN 2397-3374. doi: 10.1038/s41562-019-0801-5. URL https://doi.org/10.1038/s41562-019-0801-5.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJXMpikCZ.

Yanbang Wang, Yen-Yu Chang, Yunyu Liu, Jure Leskovec, and Pan Li. Inductive representation learning in temporal networks via causal anonymous walks. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=KYPz4YsCPj.

Chong-Yaw Wee, Pew-Thian Yap, Wenbin Li, Kevin Denny, Jeffrey N Browndyke, Guy G Potter, Kathleen A Welsh-Bohmer, Lihong Wang, and Dinggang Shen. Enriched white matter connectivity networks for accurate identification of mci patients. *Neuroimage*, 54(3):1812–1822, 2011.

Gerald Woo, Chenghao Liu, Doyen Sahoo, Akshat Kumar, and Steven Hoi. CoST: Contrastive learning of disentangled seasonal-trend representations for time series forecasting. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=PilZY3omXV2.

Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.

Huzheng Yang, James Gee, and Jianbo Shi. Memory encoding model, 2023a.

Yi Yang, Hejie Cui, and Carl Yang. \ours: Pre-train graph neural networks for brain network analysis. In *Conference on Health, Inference, and Learning*, pp. 526–544. PMLR, 2023b.

Wenchao Yu, Wei Cheng, Charu C. Aggarwal, Kai Zhang, Haifeng Chen, and Wei Wang. Netwalk: A flexible deep embedding approach for anomaly detection in dynamic networks. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pp. 2672–2681, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355520. doi: 10.1145/3219819.3220024. URL https://doi.org/10.1145/3219819.3220024.

Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. Ts2vec: Towards universal representation of time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8980–8987, 2022.

George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 2114–2124, 2021.

Ruochi Zhang, Yuesong Zou, and Jian Ma. Hyper-sagnn: a self-attention based graph neural network for hypergraphs. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryeHuJBtPH.

Yanqiao Zhu, Hejie Cui, Lifang He, Lichao Sun, and Carl Yang. Joint embedding of structural and functional brain networks with graph neural networks for mental illness diagnosis. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 272–276. IEEE, 2022.