A Generative Self-Supervised Framework using Functional Connectivity in fMRI Data

Jungwon Choi[1], **Seongho Keum**[1], **EungGu Yun*** , **Byung-Hoon Kim**[† 2 3], **Juho Lee**[† 1 4]
[1]KAIST AI, [2]Yonsei University College of Medicine, [3]MGH, Harvard Medical School, [4]AITRICS
{jungwon.choi, shkeum, eunggu.yun}@kaist.ac.kr,
egyptdj@yonsei.ac.kr, juholee@kaist.ac.kr

Abstract

Deep neural networks trained on Functional Connectivity (FC) networks extracted from functional Magnetic Resonance Imaging (fMRI) data have gained popularity due to the increasing availability of data and advances in model architectures, including Graph Neural Network (GNN). Recent research on the application of GNN to FC suggests that exploiting the time-varying properties of the FC could significantly improve the accuracy and interpretability of the model prediction. However, the high cost of acquiring high-quality fMRI data and corresponding phenotypic labels poses a hurdle to their application in real-world settings, such that a model naïvely trained in a supervised fashion can suffer from insufficient performance or a lack of generalization on a small number of data. In addition, most Self-Supervised Learning (SSL) approaches for GNNs to date adopt a *contrastive* strategy, which tends to lose appropriate semantic information when the graph structure is perturbed or does not leverage both spatial and temporal information simultaneously. In light of these challenges, we propose a *generative* SSL approach that is tailored to effectively harness spatio-temporal information within dynamic FC. Our empirical results, experimented with large-scale (>50,000) fMRI datasets, demonstrate that our approach learns valuable representations and enables the construction of accurate and robust models when fine-tuned for downstream tasks.

1 Introduction

The investigation into the complexities of human brain functionality has seen significant strides with the advent of neuro-imaging techniques [17]. Among these, fMRI is considered a pivotal modality. It captures Blood-Oxygen-Level-Dependent (BOLD) signals, offering an in-depth view of the brain's neural activity with relatively high spatial and temporal resolution. Leveraging FC based on fMRI data has become increasingly popular in solving a myriad of problems related to the human brain [2, 14]. FC allows the formation of graphs that represent connections between Regions of Interests (ROIs) in the brain, thereby transforming the problem into a graph-learning task.

To add to the complexity, acquiring labeled fMRI data is an expensive and laborious process, often resulting in limited availability of labeled data for supervised learning [20, 1]. This challenge is not unique to fMRI but is a common hurdle in many real-world applications such as fraud detection, event forecasting, and recommendation systems. SSL thus appears as a compelling solution to leverage the plethora of unlabeled fMRI data to learn useful features for downstream tasks [6, 8, 31].

However, most existing SSL approaches for graph data, including FC networks, focus solely on static graphs, ignoring the temporal dynamics that are often crucial for understanding complex systems [28, 32, 11, 25, 19, 18, 12]. This is a significant limitation, as many real-world networks, including

*Independent researcher / †Corresponding author.

Temporal Graph Learning Workshop @ NeurIPS 2023, New Orleans.

brain networks, social networks, and financial systems, are inherently dynamic. They evolve over time, and this temporal information can be crucial for various applications like anomaly detection and recommendation systems.

To address this gap, we introduce a novel framework named Spatio-Temporal Masked AutoEncoder (ST-MAE) specifically tailored for fMRI data. Unlike conventional methods that mask nodes or edges in static graphs, ST-MAE learns node representations that capture the temporal knowledge inherent in dynamic graphs. Specifically, ST-MAE employs representations from different time stamps to reconstruct masked node features at intermediate time stamps. We pre-train our model on a large-scale UKB [24] dataset, comprising approximately 40,000 entries, transforming it into FC-based dynamic graphs. Our methodology undergoes extensive validation against various benchmarks including ABCD [5], HCP [27], HCP-A [3], HCP-D [23], ABIDE [7], and ADHD200 [4]. The results demonstrate a notable improvement in downstream fMRI tasks.

The primary contributions of our work are as follows:

- We are the first to propose a generative SSL framework for dynamic graphs that takes into account temporal features for pre-training, introducing the concept of Spatio-Temporal Masked AutoEncoder (ST-MAE).
- We utilize the large-scale UKB dataset to create FC-based dynamic graphs and demonstrate the capability of SSL in capturing meaningful fMRI representations for downstream tasks.
- Our framework excels particularly in the classification of psychiatric disorders, highlighting its utility in scenarios with limited labeled data.

2 Background

2.1 Settings and Notations for Dynamic Graphs

A static graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consists of a vertex set $\mathcal{V}$ and an edge set $\mathcal{E}$. In contrast, a dynamic graph G_{dyn} is defined as a sequence of graphs $\mathcal{G}(t)$ at discrete time points t. Each $\mathcal{G}(t)$ is described by an adjacency matrix $A(t)$ and node feature vectors $\boldsymbol{x}_v(t)$ where $\boldsymbol{v} \in \mathcal{V}$. Formally, a dynamic graph G_{dyn} can be defined as:

$$G_{\text{dyn}} = \{\mathcal{G}(1), \mathcal{G}(2), \dots, \mathcal{G}(T)\}, \quad \boldsymbol{A}(t) = [a_{ij}(t)] \in \{0,1\}^{N \times N}, \tag{1}$$

where the number of nodes N is assumed to be fixed throughout time and T represents the total number of timepoints in the dynamic graph. In order to capture the temporal variations in node features, we employ a time encoding vector $\boldsymbol{\eta}(t) \in \mathbb{R}^D$, which can be generated using a sequence model such as Gated Recurrent Unit (GRU) following Kim et al. [15], where D is the size of hidden dimension. The final node feature vector at time t is then defined as $\boldsymbol{x}_v(t) = \boldsymbol{W}[\boldsymbol{e}_v \| \boldsymbol{\eta}(t)]$ where $\boldsymbol{W} \in \mathbb{R}^{D \times (N+D)}$ is a learnable matrix, $\boldsymbol{e}_v \in \mathbb{R}^{N \times N}$ is the spatial feature encoding of the node, $\boldsymbol{\eta}(t)$ is the temporal feature encoding, and $\|$ is a concatenation operation.

2.2 Masked Autoencoders in Static Graph

A Masked Autoencoder for static graphs is designed to reconstruct the original graphs from partially masked graphs. In particular, given a graph with node features represented by $\boldsymbol{X}$ and an adjacency matrix denoted as $\boldsymbol{A}$, we can apply random masking to obtain $\boldsymbol{X}_m$ and $\boldsymbol{A}_m$, encode them into a representation, and then decode the representation to reconstruct the original graph. Given a masking ratio α, the masked node features $\boldsymbol{X}_m$ are constructed by substituting the randomly selected values with zeros or learnable parameters, and the masked adjacency matrix $\boldsymbol{A}_m$ is constructed by flipping randomly chosen subset of edges. Either $\boldsymbol{X}$ or $\boldsymbol{A}$ or both can be masked before being passed to the encoder, depending on the self-supervised methodology. The masked features $\boldsymbol{X}_m$ and $\boldsymbol{A}_m$ are processed by an encoder $\mathcal{F}_{\texttt{enc}}$ (usually a GNN) to be turned into a representation $\boldsymbol{Z}$, and then the representation $\boldsymbol{Z}$ is decoded via a decoder $\mathcal{F}_{\texttt{dec}}$ to yield a reconstructed node features $\hat{\boldsymbol{X}}$:

$$\boldsymbol{Z} = \mathcal{F}_{\texttt{enc}}(\boldsymbol{X}_m, \boldsymbol{A}_m), \quad \hat{\boldsymbol{X}} = \mathcal{F}_{\texttt{dec}}(\boldsymbol{Z}). \tag{2}$$

The learning objective is to minimize the discrepancy between $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$, where the discrepancy can be Mean Squared Error (MSE), Binary Cross-Entropy (BCE), or Scaled Cosine Error (SCE). The decoder is usually constructed with Multi-Layer Perceptrons (MLPs). The adjacency matrix, based on an approach proposed in Kipf and Welling [16], can be reconstructed from the representation as $\hat{\boldsymbol{A}} = \texttt{sigmoid}(\boldsymbol{Z}\boldsymbol{Z}^\top)$. The reconstruction loss between $\boldsymbol{A}$ and $\hat{\boldsymbol{A}}$ can also be included in the loss function to train the model.

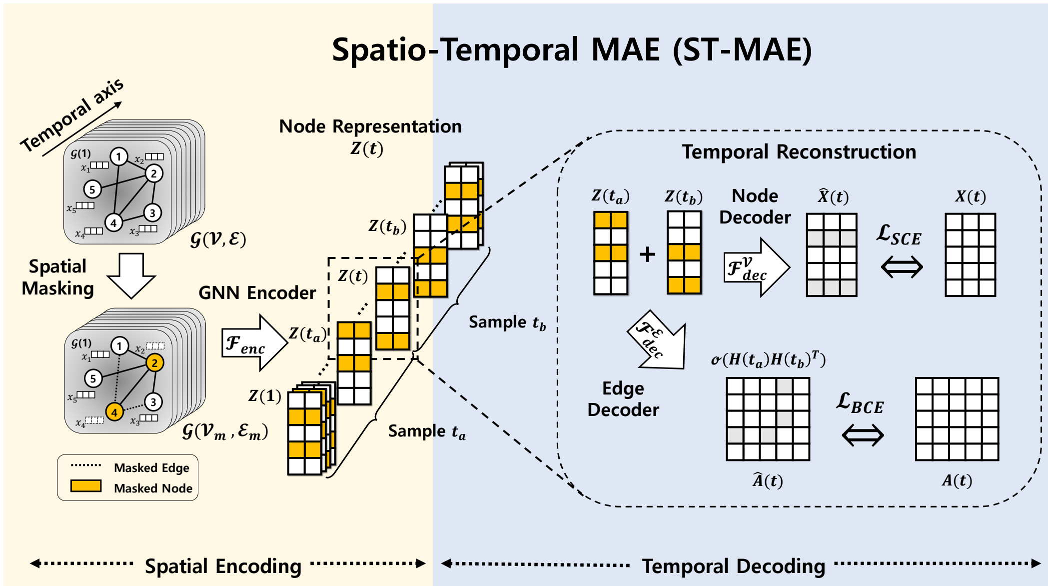

Figure 1: Spatio-Temporal Masked Autoencoder framework overview.

2.3 Constructing FC Network from fMRI Data

Following Kim et al. [15], we construct dynamic graphs out of FC networks in fMRI data by calculating the pairwise temporal correlation between the time series of different ROIs. Given a ROI-time series matrix $\boldsymbol{P} \in \mathbb{R}^{N \times T_{\text{max}}}$, the FC matrix $\boldsymbol{A}(t)$ is defined as:

$$A_{ij}(t) = \frac{\text{Cov}(p_i(t), p_j(t))}{\sigma_{p_i}(t)\sigma_{p_j}(t)} \in \mathbb{R}^{N \times N} \tag{3}$$

To transform the correlation matrix into a binary adjacency matrix, we apply thresholding to the top 30-percentile of correlation values, marking them as connected edges. All other values are treated as unconnected, as described in Kim and Ye [14].

3 ST-MAE: Spatio-temporal Masked Autoencoder Frameworks

In this study, we propose a generative SSL approach for dynamic FC of fMRI data. Unlike traditional static graph SSL methods, our approach employs a GNN encoder designed to capture knowledge in temporal graph data, as depicted in Figure 1. To facilitate this, we use a masked autoencoding objective [9] to train an encoder for spatio-temporal graphs. These encoded representations are then leveraged to perform temporal reconstruction, where nodes and edges at an intermediate timestamp are reconstructed using encodings from different time points. This enables the model to integrate and learn from both the spatial and temporal dimensions of the graph.

3.1 Masked Autoencoding Objective for Capturing Spatial Patterns

Our framework is composed of a GNN encoder $\mathcal{F}_{\texttt{enc}}$ and two decoders $\mathcal{F}^{\mathcal{V}}_{\texttt{dec}}$ and $\mathcal{F}^{\mathcal{E}}_{\texttt{dec}}$ for reconstructing node features $\boldsymbol{X}(t)$ and the adjacency matrix $\boldsymbol{A}(t)$, respectively.

We first apply the masked autoencoding objective described in the previous section for individual time-steps t. Specifically, given a time-step t, we apply the masking to the node features and the adjacency matrix $(\boldsymbol{X}(t), \boldsymbol{A}(t))$ to obtain the masked versions $(\boldsymbol{X}_{\texttt{m}}(t), \boldsymbol{A}_{\texttt{m}}(t))$, and encode them to obtain a representation $\boldsymbol{Z}(t)$.

$$\boldsymbol{Z}(t) = \mathcal{F}_{\texttt{enc}}(\boldsymbol{X}_{\texttt{m}}(t), \boldsymbol{A}_{\texttt{m}}(t)). \tag{4}$$

The node feature decoder $\mathcal{F}^{\mathcal{V}}_{\texttt{dec}}$ and the edge feature decoder $\mathcal{F}^{\mathcal{E}}_{\texttt{dec}}$ are then used to reconstruct

$$\hat{\boldsymbol{X}}(t) = \mathcal{F}^{\mathcal{V}}_{\texttt{dec}}(\boldsymbol{W}_{\texttt{single}}\boldsymbol{Z}(t)), \quad \hat{\boldsymbol{A}}(t) = \texttt{sigmoid}(\boldsymbol{H}(t)\boldsymbol{H}(t)^{\top}) \tag{5}$$

where $\boldsymbol{W}_{\texttt{single}} \in \mathbb{R}^{D \times D}$ is a learnable projection matrix and $\boldsymbol{H}(t) = \mathcal{F}^{\mathcal{E}}_{\texttt{dec}}(\boldsymbol{W}_{\texttt{single}}\boldsymbol{Z}(t))$.

Algorithm 1 Spatio-Temporal Masked AutoEncoder (ST-MAE)

Input: Dynamic graph $\mathcal{G}(t)$, Node features $\boldsymbol{X}(t)$, Edge (FC) matrix $\boldsymbol{A}(t)$
Output: Spatial encoding $\boldsymbol{Z}(t)$, Reconstructed node feature $\hat{\boldsymbol{X}}(t)$, Edge (FC) matrix $\hat{\boldsymbol{A}}(t)$
Initialize GNN encoder $\mathcal{F}_{\mathtt{enc}}$, node decoder $\mathcal{F}_{\mathtt{dec}}^{\mathcal{V}}$ and edge decoder $\mathcal{F}_{\mathtt{dec}}^{\mathcal{E}}$
for each epoch **do**
  $\mathcal{L}_{\mathtt{spatial}} \leftarrow 0$ and $\mathcal{L}_{\mathtt{temporal}} \leftarrow 0$.
  Uniformly draw a subset $\mathcal{T} \subseteq \{1, \dots, T\}$ of time-steps to apply masking.
  for $t \in \mathcal{T}$ **do**
    `/* Spatial reconstruction loss.*/`
    Mask nodes and edges in $\boldsymbol{X}(t)$ and $\boldsymbol{A}(t)$ to obtain $\boldsymbol{X}_{\mathtt{m}}(t)$ and $\boldsymbol{A}_{\mathtt{m}}(t)$
    Compute $\boldsymbol{Z}(t) \leftarrow \mathcal{F}_{\mathtt{enc}}(\boldsymbol{X}_{\mathtt{m}}(t), \boldsymbol{A}_{\mathtt{m}}(t))$.
    Compute $\hat{\boldsymbol{X}}(t) \leftarrow \mathcal{F}_{\mathtt{dec}}^{\mathcal{V}}(\boldsymbol{W}_{\mathtt{single}}\boldsymbol{Z}(t))$.
    Compute $\hat{\boldsymbol{A}}(t) = \mathtt{sigmoid}(\boldsymbol{H}(t)\boldsymbol{H}(t)^{\top})$, where $\boldsymbol{H}(t) = \mathcal{F}_{\mathtt{dec}}^{\mathcal{V}}(\boldsymbol{W}_{\mathtt{single}}\boldsymbol{Z}(t))$.
    Compute the reconstruction loss and add it to $\mathcal{L}_{\mathtt{spatial}}$.
    `/* Temporal reconstruction loss.*/`
    Uniformly sample (t_a, t_b) from $\mathcal{S}_{a,b} := \{(t_a, t_b) | 1 \leq t_a < t < t_b \leq T\}$.
    Compute $\boldsymbol{Z}(t_a)$ and $\boldsymbol{Z}(t_b)$ with $\mathcal{F}_{\mathtt{enc}}$.
    Compute $\hat{\boldsymbol{X}}_{a,b}(t) = \mathcal{F}_{\mathtt{dec}}^{\mathcal{V}}(\boldsymbol{W}_{\mathtt{concat}}[\boldsymbol{Z}(t_a) \| \boldsymbol{Z}(t_b)])$.
    Compute $\hat{\boldsymbol{A}}_{a,b}(t) = \frac{1}{2}\Big(\mathtt{sigmoid}(\boldsymbol{H}(t_a)\boldsymbol{H}(t_b)^{\top}) + \mathtt{sigmoid}(\boldsymbol{H}(t_b)\boldsymbol{H}(t_a)^{\top})\Big)$.
    Compute the reconstruction loss and add it to $\mathcal{L}_{\mathtt{temporal}}$.
  end for
  Compute the overall loss $\mathcal{L}_{\mathtt{ST\text{-}MAE}} = \mathcal{L}_{\mathtt{spatial}} + \mathcal{L}_{\mathtt{temporal}}$.
  Update the model parameters by taking the gradient descent step with $\mathcal{L}_{\mathtt{ST\text{-}MAE}}$.
end for

At each training step, based on a pre-defined masking ratio, we pick a subset $\mathcal{T} \subseteq \{1, \dots, T\}$ of time-steps and compute the reconstruction loss for those time-steps. We choose the SCE loss for the node reconstruction and the BCE loss for the adjacency reconstruction, constituting the spatial reconstruction loss,

$$\mathcal{L}_{\mathtt{spatial}} = \sum_{t \in \mathcal{T}} \Big(\mathcal{L}_{\mathtt{sce}}(\boldsymbol{X}(t), \hat{\boldsymbol{X}}(t)) + \mathcal{L}_{\mathtt{bce}}(\boldsymbol{A}(t), \hat{\boldsymbol{A}}(t))\Big). \tag{6}$$

3.2 Temporal Reconstruction Objective

To further encourage the encoder to capture the temporal dynamics in graphs, we employ the additional task for our SSL framework which is to predict a graph at a time step t based on the representations computed from the graphs at nearby time steps. More specifically, for $t \in \mathcal{T}$, we first draw two timesteps t_a and t_b uniformly from $\mathcal{S}_{a,b} := \{(t_a, t_b) | 1 \leq t_a < t < t_b \leq T\}$. The task is to reconstruct $(\hat{\boldsymbol{X}}_{a,b}(t), \hat{\boldsymbol{A}}_{a,b}(t))$ based on the representations $\boldsymbol{Z}(t_a)$ and $\boldsymbol{Z}(t_b)$, not based on the representation computed from the masked version of the graph $(\boldsymbol{X}(t), \boldsymbol{A}(t))$ as before. The node feature decoder $\mathcal{F}_{\mathtt{dec}}^{\mathcal{V}}$ reconstructs the node feature $\hat{\boldsymbol{X}}_{a,b}(t)$ based on two representations,

$$\hat{\boldsymbol{X}}_{a,b}(t) = \mathcal{F}_{\mathtt{dec}}^{\mathcal{V}}(\boldsymbol{W}_{\mathtt{concat}}[\boldsymbol{Z}(t_a) \| \boldsymbol{Z}(t_b)]) \tag{7}$$

where $\boldsymbol{W}_{\mathtt{concat}} \in \mathbb{R}^{D \times 2D}$ is a learnable projection matrix. The adjacency matrix is reconstructed similarly, but using two representations $\boldsymbol{Z}(t_a)$ and $\boldsymbol{Z}(t_b)$,

$$\boldsymbol{H}(t_a) = \mathcal{F}_{\mathtt{dec}}^{\mathcal{E}}(\boldsymbol{W}_{\mathtt{single}}\boldsymbol{Z}(t_a)),\ \boldsymbol{H}(t_b) = \mathcal{F}_{\mathtt{dec}}^{\mathcal{E}}(\boldsymbol{W}_{\mathtt{single}}\boldsymbol{Z}(t_b)), \tag{8}$$

$$\hat{\boldsymbol{A}}_{a,b}(t) = \frac{1}{2}\Big(\mathtt{sigmoid}(\boldsymbol{H}(t_a)\boldsymbol{H}(t_b)^{\top}) + \mathtt{sigmoid}(\boldsymbol{H}(t_b)\boldsymbol{H}(t_a)^{\top})\Big). \tag{9}$$

Then we compute the temporal reconstruction loss similar to the spatial reconstruction loss as,

$$\mathcal{L}_{\mathtt{temporal}} = \sum_{t \in \mathcal{T}} \Big(\mathcal{L}_{\mathtt{sce}}(\hat{\boldsymbol{X}}_{a,b}(t), \boldsymbol{X}(t)) + \mathcal{L}_{\mathtt{bce}}(\hat{\boldsymbol{A}}_{a,b}(t), \boldsymbol{A}(t))\Big). \tag{10}$$

Table 1: Statistics of dynamic graphs in fMRI datasets. The variables represent the following; $|G|$: number of graphs, $|N|$: number of nodes, $|E|$: number of edges, d_{max}: the maximum degree of nodes in each dataset, d_{avg}: average degree of nodes in each dataset, K: global clustering coefficient.

Dataset	$\|G\|$	$\|N\|_{avg}$	$\|E\|_{avg}$	d_{max}	d_{avg}	K
UKB	1,145,564	400	23,800	264	119	0.662
ABCD	191,331	400	23,800	285	119	0.663
HCP	78,696	400	23,800	281	119	0.601
HCP-A	19,548	400	23,800	287	119	0.644
HCP-D	17,064	400	23,800	262	119	0.633
ABIDE	83,096	400	23,800	278	119	0.616
ADHD200	36,126	400	23,800	257	119	0.592

3.3 Overall Training Pipeline

At each step, we compute the spatial reconstruction loss $\mathcal{L}_{\texttt{spatial}}$ and the temporal reconstruction loss $\mathcal{L}_{\texttt{temporal}}$. The overall loss function $\mathcal{L}_{\texttt{ST-MAE}}$ is defined as the sum of the two objectives.

$$\mathcal{L}_{\texttt{ST-MAE}} = \mathcal{L}_{\texttt{spatial}} + \mathcal{L}_{\texttt{temporal}} \tag{11}$$

We call our generative SSL framework based on masked autoencoder the Spatio-Temporal Masked AutoEncoder (ST-MAE) for dynamic graphs. Algorithm 1 summarizes the overall training pipeline of ST-MAE.

4 Experiments

Datasets. We compare our proposed method with several state-of-the-art SSL methods on a collection of publicly available resting-state fMRI datasets including both static and dynamic circumstances. We preprocess fMRI data into dynamic graphs with FC of 400 ROIs. As UKB [24] consists of 40,913 samples, which is one of the largest public fMRI datasets, we use it for pre-training of the GNN encoder. Then, we present downstream findings on six datasets: ABCD [5], HCP [27], HCP-A [3], HCP-D [23], ABIDE [10], and ADHD200 [4]. Graph statistics under dynamic settings are in Table 1. Please refer to the details of the datasets in Appendix A, and for information on the compared SSL methods, see Appendix B.

4.1 Experimental Details

To construct dynamic graphs, we employed a window size and stride of 50 and 16, respectively, for the UKB, ABCD, HCP, HCP-A, and HCP-D datasets. For the ABIDE and ADHD200 datasets, we used values of 16 and 3. Additionally, we followed a procedure akin to that described in Kim et al. [15], wherein each batch containing ROI-timestamps of fixed length sampled randomly per dataset.

For the baseline of our experiment, we employed a 4-layer Graph Isomorphism Network (GIN) [29] as GNN encoder. Following Kim et al. [15], to obtain the graph representation, We used SERO as the readout function and leveraged a jumping knowledge network [30] architecture, which concatenates dynamic graph representations across layers.

We evaluated the downstream performance for tasks such as gender classification and age regression on a diverse set of public fMRI datasets, including ABCD, HCP, HCP-A, HCP-D, ABIDE, and ADHD200. Furthermore, to assess potential improvements in clinical classification, we tested psychiatric disorder classification performance on the ABIDE and ADHD200 datasets. We use Adam optimizer with a learning rate of 0.0005 and a weight decay of 0.0001. During pre-training, we used a cosine decay learning rate scheduler, while for fine-tuning, a one-cycle scheduler was employed. Specifically, the learning rate increased gradually to 0.001 during the initial 20% of the training epochs and then decreased to 5.0×10^{-7}. Our approach was consistently trained with a batch size of 32. All experiments were conducted on an NVIDIA GeForce RTX 3090. The fine-tuning performance was averaged over 5-fold cross-validation.

Table 2: Results for gender classification tasks across fMRI datasets. Scores represent the Area Under the Receiver Operating Characteristics (AUROC).

Type of FC	Methods	Train Type	ABCD	HCP	HCPA	HCPD	ABIDE	ADHD200	Rank
Static	Baseline	Supervised	83.75	86.31	68.36	65.13	68.81	61.60	6.83
	DGI [26]	Contrastive SSL	73.80	87.02	70.15	68.20	67.89	62.17	6.50
	SimGRACE [28]		73.93	87.40	69.60	66.77	70.47	65.08	5.17
	GAE [16]	Generative SSL	73.45	87.31	70.66	68.39	69.90	62.88	5.50
	VGAE [16]		72.84	87.05	68.28	65.09	71.31	64.14	6.83
	GraphMAE [11]		72.79	87.77	66.87	66.42	66.98	61.48	7.83
Dynamic	Baseline	Supervised	**85.06**	93.10	84.24	73.19	73.91	72.12	1.83
	ST-DGI [21]	Contrastive SSL	83.14	92.50	82.73	70.85	72.00	65.69	3.17
	ST-MAE (Ours)	Generative SSL	83.15	**93.58**	**86.32**	**74.92**	**77.89**	**72.68**	**1.33**

Table 3: Results for age regression tasks across fMRI datasets. Scores represent the Mean Absolute Error (MAE).

Type of FC	Methods	Train Type	ABCD	HCP	HCPA	HCPD	ABIDE	ADHD200	Rank
Static	Baseline	Supervised	**0.51**	3.11	9.44	2.51	4.39	2.07	6.67
	DGI [26]	Contrastive SSL	0.54	3.12	9.38	2.50	4.27	2.03	5.33
	SimGRACE [28]		0.54	3.09	9.48	2.35	4.28	1.97	5.00
	GAE [16]	Generative SSL	0.54	3.12	9.29	2.42	4.37	2.06	6.00
	VGAE [16]		0.54	3.13	9.40	2.39	4.26	2.07	6.33
	GraphMAE [11]		0.54	3.08	9.43	2.48	4.41	2.05	6.50
Dynamic	Baseline	Supervised	0.55	**2.74**	8.39	2.16	**4.12**	1.97	3.50
	ST-DGI [21]	Contrastive SSL	0.54	2.84	**7.93**	2.15	4.18	1.93	3.00
	ST-MAE (Ours)	Generative SSL	0.54	2.82	**7.93**	**2.06**	4.13	**1.86**	**2.67**

4.2 Downstream-task Performance

We evaluated the performance of ST-MAE using multiple publicly available fMRI datasets, with particular emphasis on gender classification, age regression, and psychiatric diagnosis classification tasks. The empirical results reported in Table 2, 3, and 4 clearly show that our method consistently outperforms both self-supervised and supervised baselines across all tasks.

For gender classification in Table 2, ST-MAE achieved the highest AUROC scores, particularly excelling in dynamic FC with an AUROC improvement of 3.98% over the baseline on the ABIDE dataset. Similarly, in the age regression task in Table 3, ST-MAE demonstrated superiority by achieving the lowest MAE in the HCP-D and ADHD200 datasets. Moreover, in psychiatric diagnosis classification in Table 4, particularly where labeled data are scarce, ST-MAE outperforms other methods on the ABIDE and ADHD200 datasets.

These results validate the effectiveness of ST-MAE in capturing both spatial and temporal dynamics, while also highlighting its broad applicability and robustness in real-world scenarios. Importantly, by leveraging SSL, ST-MAE addresses the challenge of limited labeled data, making it particularly impactful for advancing research in neuropsychiatric disorders and other healthcare applications reliant on fMRI data analysis.

4.3 Ablation Study

We aimed to take full advantage of the large number of unlabeled fMRI data to develop a useful fMRI representation through SSL for downstream tasks with relatively limited data. To demonstrate the effectiveness of ST-MAE, we conducted an ablation study on the number of data for SSL and labeled data ratio for downstream task, and reconsruction strategies.

4.3.1 Effectiveness of Large-scale fMRI Datasets

We examined the impact of the amount of UKB data used for SSL on downstream performance, using gender classification on the ABIDE dataset as a case study. As shown in Figure 2, we confirmed our intuition that performance increases as the amount of data used for SSL increases. This confirms that it is possible to learn a meaningful fMRI representation from large scale fMRI data though SSL.

Table 4: Results for psychiatric diagnosis classification tasks on ABIDE and ADHD200 datasets.

Type of FC	Methods	Train Type	ABIDE Acc. (↑)	ABIDE AUROC (↑)	ADHD200 Acc. (↑)	ADHD200 AUROC (↑)	Rank
Static	Baseline	Supervised	58.94	63.78	49.47	55.74	7.00
	DGI [26]	Contrastive SSL	60.52	64.44	49.17	54.94	7.25
	SimGRACE [28]		60.97	66.14	45.88	54.50	7.00
	GAE [16]	Generative SSL	61.09	65.14	48.57	55.96	5.50
	VGAE [16]		62.44	65.04	50.67	58.50	4.00
	GraphMAE [11]		61.65	64.46	52.01	55.37	5.25
Dynamic	Baseline	Supervised	63.01	67.58	52.47	58.27	2.25
	ST-DGI [21]	Contrastive SSL	62.79	67.03	48.27	54.47	5.75
	ST-MAE (Ours)	Generative SSL	**64.48**	**69.03**	**53.07**	**59.35**	**1.00**

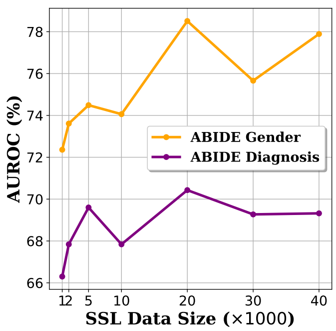

Figure 2: The effect of the number of data for SSL.

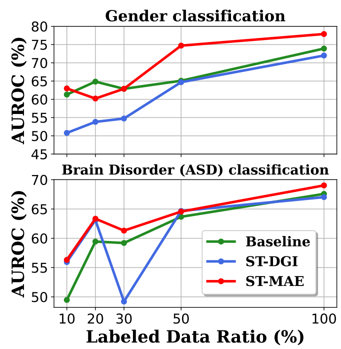

Figure 3: ABIDE classification results on limited data.

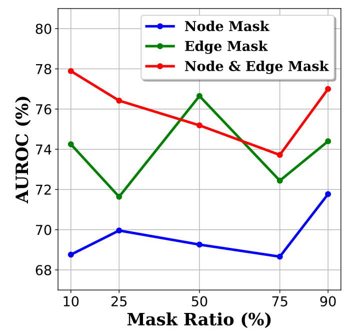

Figure 4: The ablation result of mask ratio on ABIDE dataset.

4.3.2 Effectiveness for Limited Data

In scenarios with a limited number of labels, we reduced the percentage of labeled data used for downstream training to see if ST-MAE could achieve better performance with less data. In Figure 3, we observe that the model performing SSL with ST-MAE achieves better performance even when trained using less data, suggesting that it provided a more useful starting point for downstream tasks.

4.3.3 Ablation of Masking Ratio

To examine the impact of reconstruction methods and masking ratios on downstream task performance, we experimented with different masking ratios for node and edge reconstruction and compared their gender classification performance on the ABIDE dataset. As shown in Figure 4, reconstructing both nodes and edges together is more effective than learning them separately. Furthermore, performance variations due to masking ratios mirrored those observed in the individual reconstruction targets. Since performance can vary depending on the masking ratio, it is reasonable to specify the appropriate masking ratio according to the task.

4.3.4 Ablation of Reconstruction Criterion

We compared the reconstruction criterion used in ST-MAE with different criteria for each of the node and edge reconstructions. For node reconstruction, we used MSE and SCE, and for edge reconstruction, we used MSE and BCE to compare the effectiveness of each combination. As shown in Table 5, we found the best combination when using SCE as the node restoration criterion and BCE as the edge restoration criterion, and this combination was incorporated into our ST-MAE framework.

Table 5: Ablation results of reconsruction criterion on ABIDE and ADHD200 datasets

Criterion Node	Criterion Edge	ABIDE gender	ABIDE diagnosis	ADHD200 gender	ADHD200 diagnosis
MSE	MSE	72.21	64.26	64.88	54.06
MSE	BCE	74.20	63.57	69.25	54.22
SCE	MSE	74.96	63.58	66.49	55.10
SCE	BCE	**77.89**	**65.05**	**70.29**	**55.27**

5 Related Works

5.1 Self-supervised Learning on Static Graphs

SSL on static graphs has emerged as a compelling approach to extract useful representations from graph-structured data without requiring explicit labels. These methods are generally classified into two categories: contrastive SSL and generative SSL. Both approaches aim to generate informative node and edge features that are useful for a variety of downstream tasks, such as node classification, link prediction, and graph classification.

Contrastive Self-supervised Learning Contrastive SSL techniques in graphs aim to learn embeddings by maximizing the similarity between closely related representations while minimizing the similarity between unrelated representations. DGI [26] pioneered this approach by maximizing mutual information between local patches and the entire graph. GCL [31] introduced graph augmentations to create diverse and informative positive pairs for contrastive learning. Though these methods offer better generalization capabilities, they come at the cost of computational efficiency. To mitigate this, SimGRACE [28] provided a simplified approach without complex data augmentations by utilizing correlated views obtained through a GNN with perturbed parameters. Additionally, SimGCL [32] achieved competitive performance in recommendation systems by adding uniform noise to the embedding space, generating contrastive views without the need for graph augmentations.

Generative Self-supervised Learning Generative SSL in graphs primarily focuses on reconstructing the original graph or its features from partially masked or perturbed node or edge features. VGAE [16], a pioneering work in generative SSL, proposed a method that reconstructs a graph's adjacency matrix using node representations. It employed Variational Auto Encoder (VAE) for unsupervised learning in graph-structured data, achieving effective performance in link prediction tasks. GraphMAE [11], as one of the earliest works in this area, concentrated on the reconstruction of node features and demonstrated superior performance in node and graph classification tasks over traditional contrastive self-supervised learning methods. Building on this, GraphMAE2 [12] introduced multi-view random masking and regularization, further enhancing generalization performance. However, these methods primarily focus on static graphs and do not consider learning the temporal dynamics inherent in dynamic graphs.

5.2 Self-supervised Learning on Dynamic Graphs

SSL techniques for dynamic graphs are relatively less explored, especially in the medical domain. These methods aim to capture the evolving nature of graphs, emphasizing the temporal relationships among nodes in addition to the spatial structure. Pioneering work has been conducted in non-medical domains, such as traffic flow prediction [21, 33, 13]. For instance, ST-DGI [21] has shown how generative SSL can be effective for time-series graph data, particularly in overcoming distribution shift issues commonly seen in contrastive approaches.

5.3 Deep Neural Networks on Spatio-Temporal Graphs

Deep learning on spatio-temporal graphs is a burgeoning field that aims to capture both the spatial relationships and temporal dynamics in graph-structured data. STAGIN [15] was a seminal work that successfully integrated both spatial and temporal aspects, setting a new performance benchmark across multiple tasks. This serves as our baseline for SSL on spatio-temporal graphs. Following this, Said et al. [22] introduced a graph-based neuroimaging datasets and demonstrated performance improvements by utilizing sparser graphs and a larger number of ROIs.

6 Conclusion

In this study, we presented Spatio-Temporal Masked AutoEncoder (ST-MAE), a SSL framework tailored for fMRI dynamic graphs. Our method has shown robust and superior performance in various downstream tasks, ranging from gender classification to psychiatric diagnosis classifcation. Our work contributes to both the fMRI research community and the broader field of SSL, especially in settings where labeled data are limited. The findings affirm that ST-MAE excels not only in capturing spatio-temporal dynamics but also in its adaptability for a wide range of applications. We believe this work opens up new possibilities for more advanced analytics in multiple domains.

Acknowledgments

This work was partly supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (NRF-2022R1I1A1A01069589), the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (NRF-2021M3E5D9025030) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2019-0-00075, Artificial Intelligence Graduate School Program (KAIST)).

References

[1] Fidel Alfaro-Almagro, Mark Jenkinson, Neal K Bangerter, Jesper LR Andersson, Ludovica Griffanti, Gwenaëlle Douaud, Stamatios N Sotiropoulos, Saad Jbabdi, Moises Hernandez-Fernandez, Emmanuel Vallee, et al. Image processing and quality control for the first 10,000 brain imaging datasets from uk biobank. *Neuroimage*, 166:400–424, 2018. 1

[2] Salim Arslan, Sofia Ira Ktena, Ben Glocker, and Daniel Rueckert. Graph saliency maps through spectral convolutional networks: Application to sex classification with brain connectivity. In *Graphs in Biomedical Image Analysis and Integrating Medical Imaging and Non-Imaging Modalities: Second International Workshop, GRAIL 2018 and First International Workshop, Beyond MIC 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 2*, pages 3–13. Springer, 2018. 1

[3] Susan Y Bookheimer, David H Salat, Melissa Terpstra, Beau M Ances, Deanna M Barch, Randy L Buckner, Gregory C Burgess, Sandra W Curtiss, Mirella Diaz-Santos, Jennifer Stine Elam, et al. The lifespan human connectome project in aging: an overview. *Neuroimage*, 185: 335–348, 2019. 2, 5, 12

[4] Matthew RG Brown, Gagan S Sidhu, Russell Greiner, Nasimeh Asgarian, Meysam Bastani, Peter H Silverstone, Andrew J Greenshaw, and Serdar M Dursun. Adhd-200 global competition: diagnosing adhd using personal characteristic data can outperform resting state fmri measurements. *Frontiers in systems neuroscience*, 6:69, 2012. 2, 5, 12

[5] Betty Jo Casey, Tariq Cannonier, May I Conley, Alexandra O Cohen, Deanna M Barch, Mary M Heitzeg, Mary E Soules, Theresa Teslovich, Danielle V Dellarco, Hugh Garavan, et al. The adolescent brain cognitive development (abcd) study: imaging acquisition across 21 sites. *Developmental cognitive neuroscience*, 32:43–54, 2018. 2, 5, 12

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1

[7] 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):5, 2013. 2, 12

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[9] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[10] Anibal Sólon Heinsfeld, Alexandre Rosa Franco, R Cameron Craddock, Augusto Buchweitz, and Felipe Meneguzzi. Identification of autism spectrum disorder using deep learning and the abide dataset. *NeuroImage: Clinical*, 17:16–23, 2018. 5

[11] Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. Graphmae: Self-supervised masked graph autoencoders. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 594–604, 2022. 1, 6, 7, 8, 12

[12] Zhenyu Hou, Yufei He, Yukuo Cen, Xiao Liu, Yuxiao Dong, Evgeny Kharlamov, and Jie Tang. Graphmae2: A decoding-enhanced masked self-supervised graph learner. In *Proceedings of the ACM Web Conference 2023*, pages 737–746, 2023. 1, 8, 12

[13] Jiahao Ji, Jingyuan Wang, Chao Huang, Junjie Wu, Boren Xu, Zhenhe Wu, Junbo Zhang, and Yu Zheng. Spatio-temporal self-supervised learning for traffic flow prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. 8

[14] Byung-Hoon Kim and Jong Chul Ye. Understanding graph isomorphism network for rs-fmri functional connectivity analysis. *Frontiers in neuroscience*, 14:630, 2020. 1, 3

[15] Byung-Hoon Kim, Jong Chul Ye, and Jae-Jin Kim. Learning dynamic graph representation of brain connectome with spatio-temporal attention. *Advances in Neural Information Processing Systems*, 34:4314–4327, 2021. 2, 3, 5, 8

[16] Thomas N Kipf and Max Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016. 2, 6, 7, 8, 12

[17] Sofia Ira Ktena, Sarah Parisot, Enzo Ferrante, Martin Rajchl, Matthew Lee, Ben Glocker, and Daniel Rueckert. Metric learning with spectral graph convolutions on brain connectivity networks. *NeuroImage*, 169:431–442, 2018. 1

[18] Haifeng Li, Jun Cao, Jiawei Zhu, Qinyao Luo, Silu He, and Xuying Wang. Augmentation-free graph contrastive learning of invariant-discriminative representations. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 1

[19] Jintang Li, Ruofan Wu, Wangbin Sun, Liang Chen, Sheng Tian, Liang Zhu, Changhua Meng, Zibin Zheng, and Weiqiang Wang. What's behind the mask: Understanding masked graph modeling for graph autoencoders. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1268–1279, 2023. 1

[20] Karla L Miller, Fidel Alfaro-Almagro, Neal K Bangerter, David L Thomas, Essa Yacoub, Junqian Xu, Andreas J Bartsch, Saad Jbabdi, Stamatios N Sotiropoulos, Jesper LR Andersson, et al. Multimodal population brain imaging in the uk biobank prospective epidemiological study. *Nature neuroscience*, 19(11):1523–1536, 2016. 1

[21] Felix L Opolka, Aaron Solomon, Cătălina Cangea, Petar Veličković, Pietro Liò, and R Devon Hjelm. Spatio-temporal deep graph infomax. *arXiv preprint arXiv:1904.06316*, 2019. 6, 7, 8, 12

[22] Anwar Said, Roza G Bayrak, Tyler Derr, Mudassir Shabbir, Daniel Moyer, Catie Chang, and Xenofon Koutsoukos. Neurograph: Benchmarks for graph machine learning in brain connectomics. *arXiv preprint arXiv:2306.06202*, 2023. 8

[23] Leah H Somerville, Susan Y Bookheimer, Randy L Buckner, Gregory C Burgess, Sandra W Curtiss, Mirella Dapretto, Jennifer Stine Elam, Michael S Gaffrey, Michael P Harms, Cynthia Hodge, et al. The lifespan human connectome project in development: A large-scale study of brain connectivity development in 5–21 year olds. *Neuroimage*, 183:456–468, 2018. 2, 5, 12

[24] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015. 2, 5, 12

[25] Qiaoyu Tan, Ninghao Liu, Xiao Huang, Rui Chen, Soo-Hyun Choi, and Xia Hu. Mgae: Masked autoencoders for self-supervised learning on graphs. *arXiv preprint arXiv:2201.02534*, 2022. 1

[26] Petar Veličković, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax. *arXiv preprint arXiv:1809.10341*, 2018. 6, 7, 8, 12

[27] HCP WU-Minn. 1200 subjects data release reference manual. *URL https://www. humanconnectome. org*, 565, 2017. 2, 5, 12

[28] Jun Xia, Lirong Wu, Jintao Chen, Bozhen Hu, and Stan Z Li. Simgrace: A simple framework for graph contrastive learning without data augmentation. In *Proceedings of the ACM Web Conference 2022*, pages 1070–1079, 2022. 1, 6, 7, 8, 12

[29] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018. 5

[30] Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken-ichi Kawarabayashi, and Stefanie Jegelka. Representation learning on graphs with jumping knowledge networks. In *International conference on machine learning*, pages 5453–5462. PMLR, 2018. 5

[31] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. *Advances in neural information processing systems*, 33:5812–5823, 2020. 1, 8

[32] Junliang Yu, Hongzhi Yin, Xin Xia, Tong Chen, Lizhen Cui, and Quoc Viet Hung Nguyen. Are graph augmentations necessary? simple graph contrastive learning for recommendation. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 1294–1303, 2022. 1, 8

[33] Qianru Zhang, Chao Huang, Lianghao Xia, Zheng Wang, Siu Ming Yiu, and Ruihua Han. Spatial-temporal graph learning with adversarial contrastive adaptation. In *International Conference on Machine Learning*, pages 41151–41163. PMLR, 2023. 8