# Large-scale Graph Representation Learning of Dynamic Brain Connectome with Transformers

**Byung-Hoon Kim**[*†1 2], **Jungwon Choi**[* 3], **EungGu Yun**[‡], **Kyungsang Kim**[2], **Xiang Li**[2], **Juho Lee**[† 3 4]

[1]Yonsei University College of Medicine, [2]MGH, Harvard Medical School, [3]KAIST AI, [4]AITRICS

egyptdj@yonsei.ac.kr, {jungwon.choi, eunggu.yun}@kaist.ac.kr,
{kkim24, xli60}@mgh.harvard.edu, juholee@kaist.ac.kr

## Abstract

Graph Transformers have recently been successful in various graph representation learning tasks, providing a number of advantages over message-passing Graph Neural Networks. Utilizing Graph Transformers for learning the representation of the brain functional connectivity network is also gaining interest. However, studies to date have underlooked the temporal dynamics of functional connectivity, which fluctuates over time. Here, we propose a method for learning the representation of *dynamic* functional connectivity with Graph Transformers. Specifically, we define the connectome embedding, which holds the position, structure, and time information of the functional connectivity graph, and use Transformers to learn its representation across time. We perform experiments with over 50,000 resting-state fMRI samples obtained from three datasets, which is the largest number of fMRI data used in studies by far. The experimental results show that our proposed method outperforms other competitive baselines in gender classification and age regression tasks based on the functional connectivity extracted from the fMRI data.

## 1 Introduction

Functional connectivity (FC) of the brain is defined as the level of neural co-activation across time between a pair of regions, measured by neuroimaging methods such as functional magnetic resonance imaging (fMRI) [11]. Based on evidence that the pattern of FC at rest can be linked to predicting one's phenotype, interest in learning the representation of the FC has been rapidly growing with the expectation that clinical phenotypes can also be predicted [10, 19]. Given the fact that FC can mathematically be regarded as a graph, graph neural networks (GNNs) have been a recent de facto choice for learning FC representations.

While researchers have witnessed promising results from the GNN-fMRI methods [3], there exist some limitations that come from the inherent structures of the model and the data. For example, the performance of GNN models in processing FC is limited by message-passing, vulnerable to over-smoothing and over-squashing with increasing depth, and requires simplifying FC into a basic graph, thus losing some of the original rich connectivity details [21].

Graph Transformers (GTs), a class of deep neural networks leveraging multi-head self-attention (MHSA), have recently shown success in various graph representation learning tasks, including in the context of functional connectivity (FC) analysis [18, 25, 20]. GTs address limitations of traditional Graph Neural Networks, such as over-smoothing, by adaptively learning weights between graph components without relying on message-passing. However, challenges in effectively embedding graph data for input into Transformers remain, with recent studies focusing on improving node and edge embeddings [8, 15, 26]. Notably, applications in FC analysis, such as those by Kan et al. [12] and Dong et al. [7], have demonstrated GTs' ability to encode brain graphs' structure and dynamics, offering new insights into FC from fMRI data. Yet, these approaches often overlook the temporal dynamics of FC, crucial for understanding brain function.

---

[*]Equal contribution / [†]Corresponding author / [‡]Independent researcher.
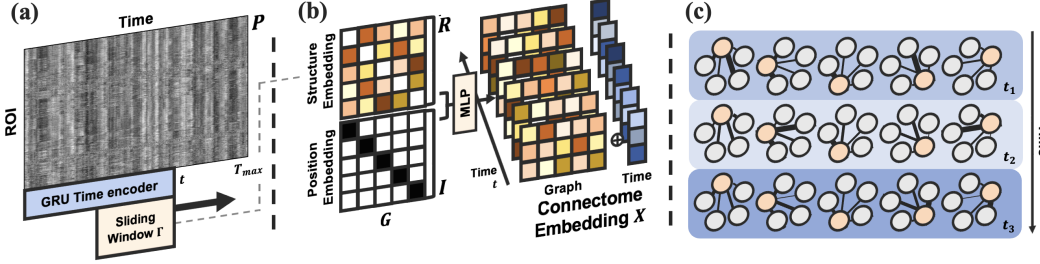
Figure 1: Defining the connectome embedding. (a) A GRU time encoder and the sliding-window dynamic FC approach are applied to the ROI-timeseries matrix. (b) Graph embedding $G$ is obtained by concatenating the structure embedding and the position embedding , followed by a feed-forward MLP. (c) The connectome embedding holds one-hop connectivity information across time at each node, which the Transformers learn self-attention weights between them.

A persistent issue in fMRI studies, including those utilizing machine learning methods such as GTs, is the challenge of replicability, with concerns about the generalizability of results to real-world data distributions [4]. While GTs have shown potential in linking fMRI signals to human phenotypes, there are lacks of evidence for their performance in external validation settings. Recent studies, however, indicate that using large-scale fMRI datasets can enhance replicability, underscoring the importance of large data volumes in fMRI research for reliable outcomes [17].

Recent studies [5, 2, 24, 1] have advanced the understanding of FC by focusing on its dynamic aspects and anomaly detection in brain networks, highlighting the importance of capturing temporal dynamics in FC. However, these studies, while advancing the field in their respective areas, often do not fully address the continuous and evolving nature of FC over time, focusing more on static or snapshot-based analysis or specific aspects like anomaly detection and generative modeling.

Here, we address these issues by training and validating a novel GT-based dynamic FC representation learning method with large-scale fMRI data. Specifically, the main goals of this work are three-fold. One is to define the connectome embedding $\boldsymbol{X}_t$ which appropriately holds the information of the brain FC at time $t$ from the raw 4D fMRI data as a combination of position, structure, and time (Section 2.1). Another is to define and train a GT $f$ such that $f : (\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_T) \to \boldsymbol{h}_{\mathrm{dyn}}$ where we input a sequence of connectome embeddings with $T$ timepoints and obtain the vector representation $\boldsymbol{h}_{\mathrm{dyn}} \in \mathbb{R}^D$, and $D$ is a pre-specified length to be output by the GT $f$ (Section 2.2). The last is to show by experiments using over 50,000 FC samples that the proposed method is capable of accurately performing classification and regression of the subject's phenotype (Section 3).

## 2  Main Contribution

### 2.1  Defining the Connectome Embedding

In defining the connectome embedding $\boldsymbol{X}_t$, we start with extracting the ROI-timeseries matrix $\boldsymbol{P}$, representing the mean BOLD signal across $N$ ROIs for $T_{\mathrm{max}}$ timepoints. The dynamic FC graph's initial position, structure, and time are encoded using a sliding-window correlation and a GRU-based time encoding approach, following [13, 14]. Specifically, the structure embedding $\boldsymbol{R}_t$ at each time $t$ is derived from the correlation coefficients within a temporal window of length $\Gamma$, shifted over time with stride $S$, forming windowed matrices $\bar{\boldsymbol{P}}_t$:

$$(\boldsymbol{R}_t)_{ij} = \frac{\mathrm{Cov}((\bar{\boldsymbol{p}}_t)_i, (\bar{\boldsymbol{p}}_t)_j)}{\sigma_{(\bar{\boldsymbol{p}}_t)_i} \sigma_{(\bar{\boldsymbol{p}}_t)_j}} \in \mathbb{R}^{N \times N},$$

where $(\boldsymbol{R}_t)ij$ captures the edge weight between nodes $i$ and $j$. The node position is separately embedded by subtracting the identity matrix from $\boldsymbol{R}_t$ to remove self-loops and then concatenating it with the identity matrix, forming $\boldsymbol{G} := [\, \boldsymbol{R}_t - \boldsymbol{I} \mid \boldsymbol{I} \,] \in \mathbb{R}^{N \times 2N}$. This graph embedding is processed through a two-layer MLP to produce a final graph embedding in $N \times D$ dimensions.

The time embedding $\eta(t) \in \mathbb{R}^D$ is the GRU output using ROI-timeseries up to the last timepoint of $\Gamma$. The final connectome embedding $\boldsymbol{X}_t$ is obtained by concatenating the MLP graph embedding with the time embedding:

$$\boldsymbol{X}_t = [\, \mathrm{MLP}(\boldsymbol{G}) \mid \eta(t) \,] \in \mathbb{R}^{(N+1) \times D}.$$

This process effectively captures the one-hop connectivity information across time at each node in the FC graph.

Table 1: Summary of the experiment datasets

| Dataset | UKB | ABCD | HCP-YA | HCP-D | HCP-A |
|---|---|---|---|---|---|
| No. Subjects | 40913 | 9111 | 1093 | 632 | 723 |
| Gender (F/M) | 21682 / 19231 | 4370 / 4741 | 594 / 499 | 339 / 293 | 405 / 318 |
| Age (Min-Max) | 40.0-70.0 | 8.9-11.0 | 22.0-37.0 | 8.1-21.9 | 36.0-89.8 |

## 2.2   TENET: Temporal Neural Transformer

In this section, we introduce the details of our proposed method, TENET, (Figure 2). To bring the learning process of the temporal information straightforwardly, we formulate the proposed method as a two-step composition of the Transformer encoders across space and time:

$$g : (\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_T) \to (\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_T),$$
$$h : (\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_T) \to \boldsymbol{h}_{\text{dyn}}$$

where $\boldsymbol{h}_t$ is a self-attended connectome feature vecture at time $t$. It can be thought that the $g$ is a self-attention across space that extracts appropriate representation at a specific timepoint, and $h$ is a self-attention across time to learn the dynamic pattern of the input fMRI signal, letting $f = h \circ g$.



Figure 2: Schematic illustration of TENET.

As mentioned above, both $g$ and $h$ incorporate the self-attention scheme to learn the relationship between each input token within the vector-stacked matrix $\boldsymbol{H}$ defined as:

$$\text{attention}(\boldsymbol{H}) = \text{softmax}\Big(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{D}}\Big)\boldsymbol{V}, \quad \boldsymbol{K} = \boldsymbol{W}_{\text{key}}\boldsymbol{H}, \ \boldsymbol{Q} = \boldsymbol{W}_{\text{query}}\boldsymbol{H}, \ \boldsymbol{V} = \boldsymbol{W}_{\text{value}}\boldsymbol{H},$$

where $\boldsymbol{K}, \boldsymbol{Q}, \boldsymbol{V}$ are transformations of input encoding to corresponding key, query, and value with learnable linear weight matrices, and $D$ is the hidden dimension. The MHSA is the self-attention parallelly projected with the multiple number of heads.

The Connectome Transformer layer processes the connectome embedding at layer $l$ using MHSA, supplemented with 1-hop connectivity $\bar{\boldsymbol{R}}_t$ and degree information, then passed through an MLP for the next-layer embedding: $\boldsymbol{Z}_t^l = \text{concatenate}(\{\text{attention}(\boldsymbol{H}_t^l), \bar{\boldsymbol{R}}_t, \Sigma^i(\bar{\boldsymbol{R}}_t)_{ij}\}), \quad \boldsymbol{H}_t^{l+1} = \text{MLP}(\boldsymbol{Z}_t^l)$. This design injects functional connectivity details into the Transformer, enhancing the model's depth and performance. The first layer embedding combines the connectome embedding with a random-initialized learnable token vector $\boldsymbol{h}_{\text{token}}$, defined as $\boldsymbol{H}_t^0 := [\boldsymbol{X}_t || \boldsymbol{h}_{\text{token}}]$. After processing through $L$ layers, the $\boldsymbol{h}_{\text{token}}$ at each timepoint $t$ represents connectome features across time. These features are further refined by $L$ layers of a standard Transformer Encoder, culminating in a final token vector used for classification or regression tasks. For a detailed exposition of the computational process, please refer to Appendix A.

# 3   Experiments

## 3.1   Dataset and Experimental Setup

We utilized three large-scale resting-state fMRI datasets: 1) UK Biobank (UKB) [16], 2) Adolescent Brain Cognitive Development (ABCD) [6], and 3) Human Connectome Project (HCP) [9], each with distinct participant age groups and preprocessing protocols. For the ABCD dataset, lacking an official preprocessed version, we employed the ABCD-HCP pipeline [2]. The ROI-timeseries matrix was extracted using the Schaefer atlas with 400 ROIs [22]. We focused on the first session of fMRI acquisition for each subject from these datasets, totaling over 50,000 samples, to mitigate sample correlation. Our experiments targeted gender classification and age regression tasks, using participant
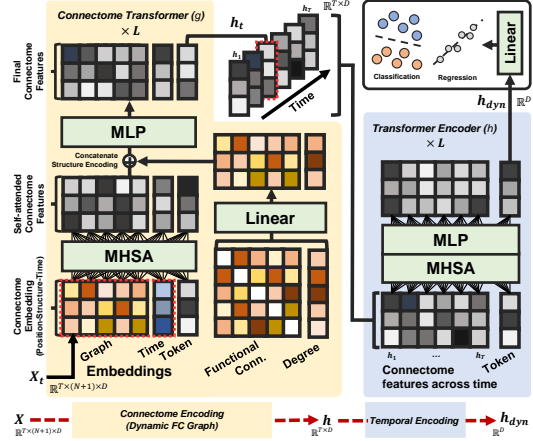
---

[2]https://github.com/DCAN-Labs/abcd-hcp-pipeline

Table 2: Performance table on our benchmark datasets.

| Dataset | HCP-YA | HCP-D | | HCP-A | | UKB | | ABCD |
|---|---|---|---|---|---|---|---|---|
| Feature | Gender | Gender | Age | Gender | Age | Gender | Age | Gender |
| TENET | 95.07 | **82.89** | **0.6878** | **89.05** | **0.6663** | 98.37 | **0.4768** | **90.21** |
| BNT | 95.49 | 81.29 | 0.6756 | 86.63 | 0.5828 | **99.04** | 0.4635 | 85.98 |
| STAGIN | **95.50** | 70.36 | 0.3966 | 84.57 | 0.4473 | 98.61 | 0.4047 | 81.61 |
| GIN | 86.49 | 62.86 | 0.3054 | 68.91 | 0.3130 | 96.67 | 0.3904 | 73.19 |

\* Performance of gender classification and age regression are reported with AUROC (%) and $R^2$ scores, respectively.

demographic data as labels. However, age regression was not applied to HCP-YA and ABCD due to limited age variability. The experiment datasets are summarized in the Table 1. It should be noted that the datasets include over 50,000 samples in total, which is a number unprecedented in any resting-state fMRI studies by far.

Our model was structured with 4 layers ($L = 4$) with each layer having a hidden dimension of 1024. The Adam optimizer, coupled with a one-cycle learning rate schedule, was used for optimization [23]. We conducted a grid search to identify the best hyperparameters, exploring batch sizes within $\{2, 4, 6, 8, 10\}$ and learning rates within $\{5 \cdot 10^{-4}, 10^{-5}, 5 \cdot 10^{-6}, 10^{-6}, 5 \cdot 10^{-7}, 10^{-7}\}$. Training involved 15 epochs for ABCD and UKB datasets and 30 epochs for HCP subsets, using a 5-fold cross-validation method. All experiments were executed on an NVIDIA GeForce RTX 3090.

### 3.2 Comparative Experiment

The performance of TENET is validated by comparing it with several baseline methods on our benchmark datasets. The baseline methods include the BNT [12], a GT-based static FC method, STAGIN [14], a GNN-based dynamic FC method, and GIN [13], a GNN-based static FC method. Performance of gender classification and age regression is evaluated with the area under the receiver operating curve (AUROC) and the $R^2$ scores, respectively.

The main results are summarized in the Table 2. From the comparative experiments, it can be seen that TENEToutperforms other baseline methods in most of the phenotype prediction tasks. We expect that this gain in performance comes from the ability of TENETto exploit dynamic information of the FC that changes over time.

### 3.3 Ablation Study

Table 3: Ablation results evaluating the impact of temporal information.

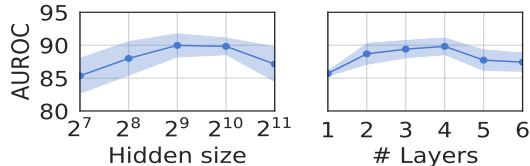| Use time encoding | Dynamic graph | AUROC |
|---|---|---|
| ✓ | ✓ | 89.82 |
| ✗ | ✓ | 88.45 |
| ✓ | ✗ | 84.57 |



Figure 3: Ablation results on model size of TENET.

We conducted an ablation study on TENET to assess the impact of temporal information, which involved two scenarios: 1) removing the GRU-derived time embedding and 2) replacing the dynamic graph embedding with a static one. The results, as detailed in Table 3, indicate a decrease in performance for both scenarios in the HCP-A gender classification task, demonstrating the importance of temporal information in our model. Additionally, sensitivity tests for the hidden dimension size and the number of layers revealed optimal performance at specific thresholds, with diminishing returns beyond these points as shown in Figure 3.

## 4 Conclusion

We propose TENET, a GT-based method for learning dynamic FC of the brain with position, structure, and time embeddings. Experiments with large-scale resting-state fMRI datasets confirm the validity of TENET. Further studies of TENETon attention interpretability, external validation performance, and theoretical understanding are expected to provide valuable insight into applying GT to the time-varying FC graph.

## Acknowledgments

## References

[1] A. Behrouz and M. Seltzer. Anomaly detection in multiplex dynamic networks: from blockchain security to brain disease prediction. *arXiv preprint arXiv:2211.08378*, 2022.

[2] A. Behrouz and M. 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)*, 2023.

[3] A. Bessadok, M. A. Mahjoub, and I. Rekik. Graph neural networks in network neuroscience. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5833–5848, 2022.

[4] R. Botvinik-Nezer and T. D. Wager. Reproducibility in neuroimaging analysis: Challenges and solutions. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 2022.

[5] A. Campbell, A. G. Zippo, L. Passamonti, N. Toschi, and P. Lio. Dyndepnet: Learning time-varying dependency structures from fmri data via dynamic graph structure learning. In *ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH)*, 2023.

[6] B. J. Casey, T. Cannonier, M. I. Conley, A. O. Cohen, D. M. Barch, M. M. Heitzeg, M. E. Soules, T. Teslovich, D. V. Dellarco, H. Garavan, et al. The adolescent brain cognitive development (abcd) study: imaging acquisition across 21 sites. *Developmental cognitive neuroscience*, 32: 43–54, 2018.

[7] Z. Dong, Y. Wu, Y. Xiao, J. S. X. Chong, Y. Jin, and J. H. Zhou. Beyond the snapshot: Brain tokenized graph transformer for longitudinal brain functional connectome embedding. *arXiv preprint arXiv:2307.00858*, 2023.

[8] V. P. Dwivedi and X. Bresson. A generalization of transformer networks to graphs. *arXiv preprint arXiv:2012.09699*, 2020.

[9] M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, J. Xu, S. Jbabdi, M. Webster, J. R. Polimeni, et al. The minimal preprocessing pipelines for the human connectome project. *Neuroimage*, 80:105–124, 2013.

[10] C. Horien, D. L. Floris, A. S. Greene, S. Noble, M. Rolison, L. Tejavibulya, D. O'Connor, J. C. McPartland, D. Scheinost, K. Chawarska, et al. Functional connectome–based predictive modeling in autism. *Biological psychiatry*, 92(8):626–642, 2022.

[11] S. A. Huettel, A. W. Song, and G. McCarthy. *Functional magnetic resonance imaging*, volume 1. Sinauer Associates Sunderland, MA, 2004.

[12] X. Kan, W. Dai, H. Cui, Z. Zhang, Y. Guo, and C. Yang. Brain network transformer. *Advances in Neural Information Processing Systems*, 35:25586–25599, 2022.

[13] B.-H. Kim and J. C. Ye. Understanding graph isomorphism network for rs-fmri functional connectivity analysis. *Frontiers in neuroscience*, 14:630, 2020.

[14] B.-H. Kim, J. C. Ye, and J.-J. Kim. Learning dynamic graph representation of brain connectome with spatio-temporal attention. *Advances in Neural Information Processing Systems*, 34:4314–4327, 2021.

[15] D. Kreuzer, D. Beaini, W. Hamilton, V. Létourneau, and P. Tossou. Rethinking graph transformers with spectral attention. *Advances in Neural Information Processing Systems*, 34: 21618–21629, 2021.

[16] T. J. Littlejohns, J. Holliday, L. M. Gibson, S. Garratt, N. Oesingmann, F. Alfaro-Almagro, J. D. Bell, C. Boultwood, R. Collins, M. C. Conroy, et al. The uk biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. *Nature communications*, 11(1):2624, 2020.

[17] S. Marek, B. Tervo-Clemmens, F. J. Calabro, D. F. Montez, B. P. Kay, A. S. Hatoum, M. R. Donohue, W. Foran, R. L. Miller, T. J. Hendrickson, et al. Reproducible brain-wide association studies require thousands of individuals. *Nature*, 603(7902):654–660, 2022.

[18] E. Min, R. Chen, Y. Bian, T. Xu, K. Zhao, W. Huang, P. Zhao, J. Huang, S. Ananiadou, and Y. Rong. Transformer for graphs: An overview from architecture perspective. *arXiv preprint arXiv:2202.08455*, 2022.

[19] E. L. Morris, S. F. Taylor, and J. Kang. On predictability of individual functional connectivity networks from clinical characteristics. *Human Brain Mapping*, 43(17):5250–5265, 2022.

[20] L. Müller, M. Galkin, C. Morris, and L. Rampášek. Attending to graph transformers. *arXiv preprint arXiv:2302.04181*, 2023.

[21] T. K. Rusch, M. M. Bronstein, and S. Mishra. A survey on oversmoothing in graph neural networks. *arXiv preprint arxiv:2303.10993*, 2023.

[22] A. Schaefer, R. Kong, E. M. Gordon, T. O. Laumann, X.-N. Zuo, A. J. Holmes, S. B. Eickhoff, and B. T. Yeo. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri. *Cerebral Cortex*, 28(9):3095–3114, 2017.

[23] L. N. Smith and N. Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, volume 11006, page 1100612. International Society for Optics and Photonics, 2019.

[24] S. E. Spasov, A. Campbell, N. Toschi, and P. Lio. Neuroevolve: A dynamic brain graph deep generative model. In *ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH)*, 2023.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[26] C. Ying, T. Cai, S. Luo, S. Zheng, G. Ke, D. He, Y. Shen, and T.-Y. Liu. Do transformers really perform badly for graph representation? *Advances in Neural Information Processing Systems*, 34:28877–28888, 2021.