Large Connectome Model: An fMRI Foundation Model of Brain Connectomes Empowered by Brain-Environment Interaction in Multitask Learning Landscape

Ziquan Wei[1], **Tingting Dan**[2], **Guorong Wu**[2,1,3*]

[1]Department of Computer Science, University of North Carolina at Chapel Hill
[2]Department of Psychiatry, University of North Carolina at Chapel Hill
[3]Neuroscience Center, University of North Carolina at Chapel Hill
ziquanw@email.unc.edu;{Tingting_Dan,grwu}@med.unc.edu

Abstract

A reliable foundation model of functional neuroimages is critical to promote clinical applications where the performance of current AI models is significantly impeded by a limited sample size. To that end, tremendous efforts have been made to pretraining large models on extensive unlabeled fMRI data using scalable self-supervised learning. Since self-supervision is not necessarily aligned with the brain-to-outcome relationship, most foundation models are suboptimal to the downstream task, such as predicting disease outcomes. By capitalizing on rich environmental variables and demographic data along with an unprecedented amount of functional neuroimages, we form the brain modeling as a multitask learning and present a scalable model architecture for (i) multitask pretraining by tokenizing multiple brain-environment interactions (BEI) and (ii) semi-supervised finetuning by assigning pseudo-labels of pretrained BEI. We have evaluated our foundation model on a variety of applications, including sex prediction, human behavior recognition, and disease early diagnosis of Autism, Parkinson's disease, Alzheimer's disease, and Schizophrenia, where promising results indicate the great potential to facilitate current neuroimaging applications in clinical routines.

Code — https://github.com/Chrisa142857/brain_network_decoder

Extended version — https://arxiv.org/abs/2510.18910

1 Introduction

A scalable foundation model dedicated to brain activity is critical to discovering the enigma of human cognition and promoting clinical applications from large-scale neuroimaging data. The topic of the brain foundation model is under exploration since BrainLM (Ortega Caro et al. 2023) via masked autoencoder. Previous works formulate this problem by mimicking natural language or image foundation models as self-regression or masking strategies, that is learning the raw signal reconstruction, e.g., Masked Autoencoder (MAE) in (Ortega Caro et al. 2023), Joint-Embedding Predictive Architecture (JEPA) in (Dong et al. 2024), and other masking methods (Wen et al. 2023; Yang et al. 2024). However, as revealed by Meta-matching (He et al. 2022), the brain connectomes share similar features across arbitrary phenotypic traits (hereafter shortened to 'phenotypes') such as age and lifestyle. The multitask learning for phenotypic prediction can hence act as the foundation objective related to neuroscience interests. Furthermore, although existing brain foundation models (Ortega Caro et al. 2023; Dong et al. 2024) take the raw signal as input, the vast majority of related works (Cui et al. 2022; Said et al. 2023; Ding et al. 2024; Wei et al. 2024) suggest that brain connectomes as the input makes more accurate predictions for clinical applications.

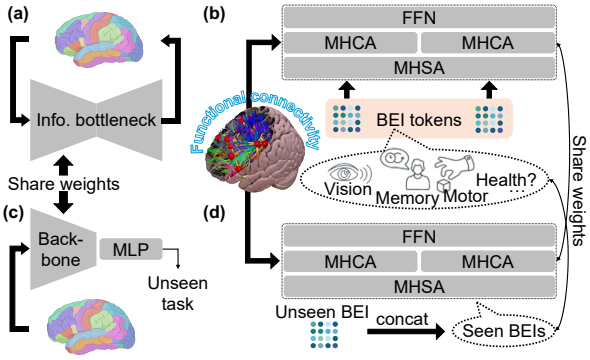

Figure 1: Learning strategies of previous brain foundation models and LCM. The pretraining in **(a)** previous brain foundation models is a reconstructive representation learning based on the information (info.) bottleneck, while **(b)** in LCM it is a multitask learning for multiple brain-environment interactions (BEI) token embeddings by a Transformer decoder, where MHSA is multi-head self-attention, MHCA is multi-head cross-attention, and FFN is a feedforward network. The finetuning in **(c)** previous studies is training a relatively small head, e.g., a multilayer perceptron (MLP), for the downstream task. **(d)** LCM finetunes the BEI tokens along with new tokens representing the downstream task.

The self-regressive methodology, although it demonstrated excellent applications such as GPTs for natural language (Achiam et al. 2023), should not be the only choice for brain foundation models. As shown in Fig. 1a, the purpose of previous brain foundation models is the same as the image/language to reconstruct the raw signal from its masked version via a bottleneck or transformer encoder architecture. Unlike image/language that is not naturally labeled, brain

*Corresponding author.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

fMRI has demographics and phenotypes, e.g., age, biological sex, cognitive state, etc, which are commonly recorded during data acquisition. However, the challenging heterogeneity in multi-phenotype learning risks the robustness of current brain foundation models due to the predictive head is relatively lightweight (Fig. 1c). According to the nature of fMRI, which always has non-imaging records, a scalable architecture of the brain foundation model is necessary to involve multitask learning from rich environmental information relevant to brain cognition.

Demographics and phenotypes that generate a diverse range of brain-environment interactions (BEI) have been found inter-correlated to each other, given the brain connectomes (He et al. 2022). Even without finetuning, the classical regression model can outperform the trained version after a basic meta-matching between phenotypes. This motivates us to develop and release a brain foundation model powered by BEI multitask learning, as illustrated in Fig. 1b. Given that cognition relevant BEIs represent the brain function used for diagnosing and other downstream tasks, the downstream outcome can be decoded from the BEI token embeddings as shown in Fig. 1d, where the cross-attention is computed between brain connectome feature and the tokenized BEI. This utilizes the findings in (He et al. 2022) that seen and unseen cognition relevant phenotypes can be meta-matched. The self-attention in Fig. 1 communicates information learned from pretrained BEIs in downstream tasks, enhancing the generality and robustness of the model.

To this end, this work presents three main contributions. (1) A new brain foundation model architecture is proposed to cooperate with multitask pretraining and semi-supervised finetuning on rich BEIs. (2) The *largest* connectome model (LCM) for brain fMRI is designed and released along with the pretrained weights based on large scale data (n =10,036). (3) Experiments on 8 fMRI datasets evaluate the performance of LCM on sex prediction, human behavior recognition, and disease early diagnosis of Autism, Parkinson's disease, Alzheimer's disease, and Schizophrenia, in terms of scalability, pretraining, and finetuning.

2 Preliminaries

The dynamic signal of brain functional MRI is a blood-oxygen dependent level (BOLD). BOLD signal, which is influenced by a mixture of factors and distorted by non-neuronal fluctuations, has a relatively low signal-to-noise ratio (SNR) (Caballero-Gaudes and Reynolds 2017). Brain connectomes, on the other hand, increase the SNR in raw signals by representing brain activity via the Pearson correlation coefficient, which is also called functional connectivity (FC).

Various works have shown superior performance using FC compared to the raw BOLD signal for downstream applications. For example, benchmark papers (Cui et al. 2022; Said et al. 2023; Ding et al. 2024) evaluated the performance by using the BOLD or the correlation as the input, and the BOLD signal has consistently demonstrated lower accuracy. In addition, both static FC and dynamic FC (using the sliding window technique) outperform the BOLD signal (Wei et al. 2024).

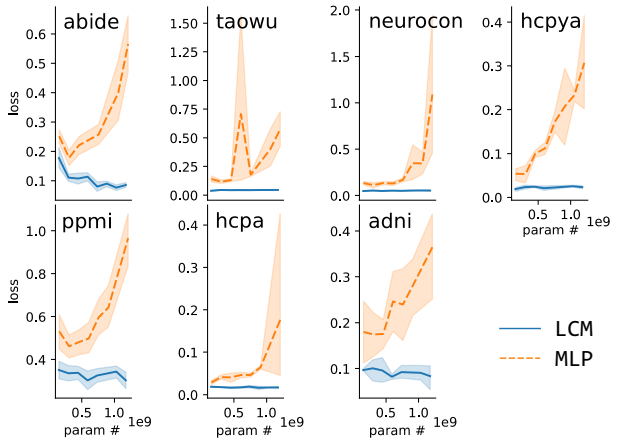

Figure 2: Scalability is demonstrated by model size vs. training loss, where the training is supervised by arbitrary non-brain-imaging phenotypes as BEIs in our multitask learning.

Therefore, we formulate the problem of the brain foundation model as a pretrained large model dedicated to the brain connectome and coined as the large connectome model (LCM). Inspired by (He et al. 2022), LCM is supervised by phenotypic labels, aka. BEI, via multitask learning, which mostly are categorical labels. However, even tremendous efforts have been made for brain connectome and graph classification or regression (Ying et al. 2021; Kan et al. 2022; Chen et al. 2023; Bedel et al. 2023; Wei et al. 2024), they, as the encoder, have a limited scalability (Wei et al. 2024). Plenty of theoretical (Keriven 2022) and experimental analysis (Rusch, Bronstein, and Mishra 2023) suggest that the reason for that includes over-smoothing and over-squashing.

Scalability analysis: For brain foundation models, MLP is commonly used as the predictive head (Ortega Caro et al. 2023; Yang et al. 2024; Dong et al. 2024). However, as shown in Fig. 2 orange curves, MLP has an exploded training loss when scaling up the parameter amount, where the MLP model is constructed with ReLU activations and residual connections in-between blocks. In this case, we need to think out of the box about the predictive head that has unsatisfactory scalability for the multitask pretraining.

Recently, (Paul et al. 2024) proposed a simple framework by replacing the MLP with the Transformer decoder as an interpretable predictive head showing similar or better performance for computer vision tasks. Based on this decoder classifier, our LCM is designed as a decoder-only architecture. In comparison, as shown in Fig. 2 blue curves, LCM shows a better scalability on seven different datasets, where training loss can be lower with more layers used in LCM. This enables LCM to learn from large-scale brain connectome data powered by BEI multitask learning. Consequently, the model efficiency surpasses previous foundation models, which are derived from vision-based encoders, as shown in Fig. 3.

Related works: BrainLM (Ortega Caro et al. 2023), to our best knowledge, is the first brain foundation model by applying MAE on BOLD signals. It fills every bit of the

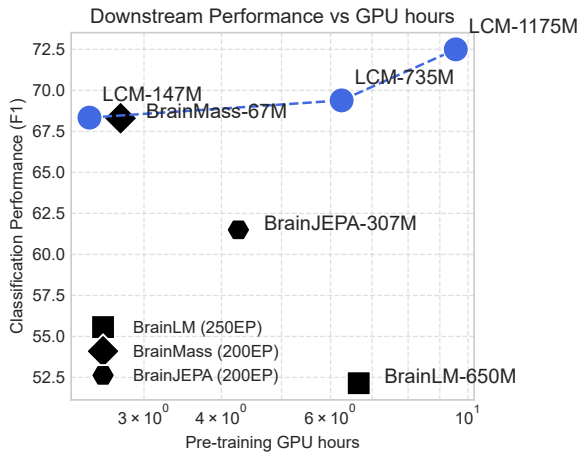

Figure 3: LCM surpasses other foundation models, demonstrating outstanding efficiency, on our biggest downstream application, ABIDE (n=1,025), as an example. Even the smallest LCM (147M), achieves comparable performance while being efficient in both parameters and resource usage.

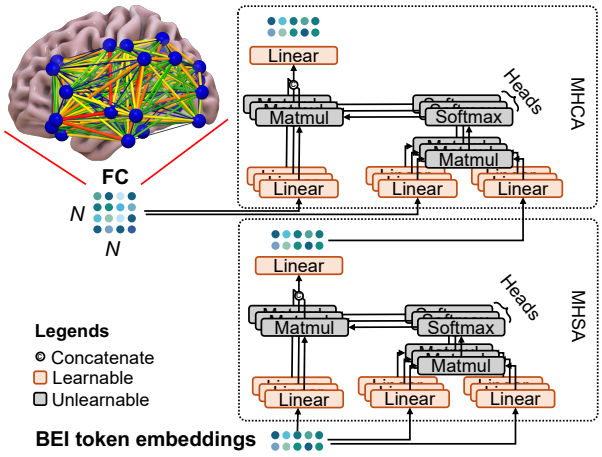

Figure 4: The architecture of one layer of the LCM.

fMRI time series can hinder the model's ability to distinguish between noise and actual signals. However, research (Assran et al. 2023) has shown that masked pretraining in generative architectures like MAE often results in suboptimal performance in off-the-shelf evaluations (e.g., linear probing). BrainJEPA (Dong et al. 2024), similarly, framed a new architecture with a different masking strategy JEPA. It handles the suboptimal issues of BrainLM by following the idea of the I-JEPA (Assran et al. 2023). Although their results have shown better performance than linear probing, none of the explicit designs have been added for learning from inter-correlated phenotypes. BrainMass (Yang et al. 2024) used a matching objective between pseudo FC matrices by masking BOLD signals. Whilst, it overlooked the phenotypes and demographics that is always assigned with the functional neuroimaging data. To this end, we present a scalable large connectome model explicitly supervised by rich environmental variables and demographic data, along with an unprecedented amount of functional neuroimages.

3 Methods

Given the FC matrix of fMRI data as defined by Sec. 2, LCM takes FC as input and learns from the supervision of multiple non-imaging records of fMRI.

Architecture

Inspired by the Transformer-based large models (Achiam et al. 2023) and (Paul et al. 2024), a decoder-only architecture as shown in Fig. 4 is employed in our LCM. The token vectors in Fig. 4 refer to N_{BEI} available BEIs in pretraining datasets, which are initialized randomly as the token embedding, denoted by $\mathbf{V} \in \mathbb{R}^{P\times E}$, where $P = \sum_i^{N_{BEI}} \mathbf{N}_i^{class}$ and $\mathbf{N}_i^{class}$ is the class number of the i^{th} categorical BEI or 1 if the BEI is a continous value. Next, $\mathbf{V}$ is updated by self-attention as follows:

$$\mathbf{V} = \texttt{Softmax}\left((\mathbf{V}\bar{\boldsymbol{\alpha}}_h)(\mathbf{V}\bar{\boldsymbol{\beta}}_h)^T/\sqrt{D}\right)(\mathbf{V}\bar{\boldsymbol{\gamma}}_h), \quad (1)$$

where $\bar{\boldsymbol{\alpha}}_h, \bar{\boldsymbol{\beta}}_h, \bar{\boldsymbol{\gamma}}_h \in \mathbb{R}^{E\times D}$ are learnable parameters of self-attention linear layers shown in Fig. 4, h is the head index, and D is the hidden channel.

Suppose $\mathbf{M} \in \mathbb{R}^{N\times N}$ is FC matrix. Cross-attention between $\mathbf{V}$ and $\mathbf{M}$ is then defined as follows

$$\mathbf{V} = \texttt{Softmax}\left((\mathbf{M}\hat{\boldsymbol{\alpha}}_h)(\mathbf{V}\hat{\boldsymbol{\beta}}_h)^T/\sqrt{D}\right)^T(\mathbf{M}\hat{\boldsymbol{\gamma}}_h), \quad (2)$$

where $\hat{\boldsymbol{\alpha}}_h, \hat{\boldsymbol{\gamma}}_h \in \mathbb{R}^{N\times D}, \hat{\boldsymbol{\beta}}_h \in \mathbb{R}^{E\times D}$ are learnable parameters of cross-attention linear layers shown in Fig. 4, and D is the hidden channel. Note that the bias in linear layers is omitted in this section for clarity.

This design allows LCM to be easily stacked since each layer updates $\mathbf{V}$ without changing the tensor shape.

Multitask Pretrain and Semi-supervised Finetune

Take categorical BEI as an example, the multitask pretraining is accomplished by

$$L_{cls} = \sum_{i=0}^{N_{BEI}} CELoss(\mathbf{V}_{S_i:S_{i+1}}, GT_i), \quad (3)$$

where $S_0 = 0, S_i = \sum_{j=0}^{i-1} \mathbf{N}_j^{class}$ if $i > 0$ and $CELoss$ denotes cross-entropy loss. The Mean Squared Error (MSE Loss) is used for regressive BEI tokens.

Finetuning objectives are the same as pretraining. Given unseen datasets that have $\hat{N}_{BEI}$ tasks as new BEIs with $\hat{\mathbf{N}}^{class}$ as the vector of class number for each BEI, the finetuning can be easily achieved by concatenating new tokens by updating $\mathbf{V} \leftarrow [\mathbf{V}, \hat{\mathbf{V}}]$, $\mathbf{N}^{class} \leftarrow [\mathbf{N}^{class}, \hat{\mathbf{N}}^{class}]$. For pretrained tokens during finetuning, the pseudo-label is assigned to them with corresponding neuroimaging configurations, e.g., 'resting-state' for clinical applications.

In case that BEIs differ on the complexity of feature representation, e.g., Parkinson's disease at different stages of

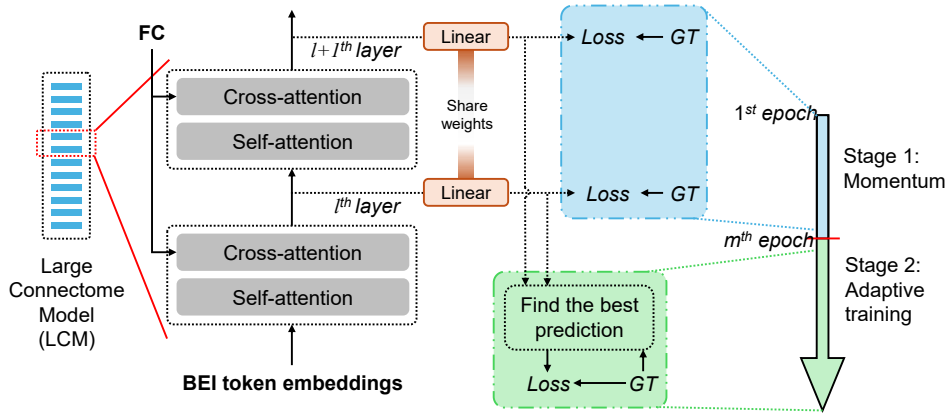

Figure 5: Pre-training and finetuning of LCM use a two-stage learning strategy: (1) Getting momentum by computing loss for all layers, and (2) adaptive training for the best layer. Note that ground truth (GT) could be a pseudo-label of the BEI, e.g., subjects are healthy by default in HCP datasets.

treatment might show different symptoms, we propose that LCM predicts each BEI at different layers of the model. As shown in Fig. 5, the BEI token embeddings from every LCM layer are stored at first. After the computation of LCM, token embeddings from all layers are input into a linear layer: $\mathbb{R}^{P\times D} \rightarrow \mathbb{R}^{P\times 1}$ to get various predictions from different layers of LCM. The best prediction, i.e., the layer that has the best predictive score, can be found given the ground truth (GT) during training. The best prediction is then used as the output to get the loss with GT. During testing, the output of the layer that has the highest score is the final output of LCM.

The initialization of our learning strategy is important to have a proper starting point and a correct direction. To account for this, as shown in Fig. 5 left part, training the LCM has two stages for both pretraining and finetuning, (1) utilize the average prediction from all layers to update the LCM parameters in the first m epochs, and (2) supervise only the best prediction in the rest of epochs. Namely, stage 1 produces a 'momentum' that can push the training of LCM to the correct direction, and hence LCM can achieve a diverse and correct feature representation for different phenotypes in the following stage.

4 Experiments

We evaluate the proposed LCM on 8 datasets including HCP Aging (HCPA), HCP Young Adult (HCPYA), ADNI, PPMI, ABIDE, Taowu, Neurocon, and SZ. Two HCPs contain more than 10,000 scans of brain fMRI from about 1,800 subjects under various cognitive states, depending on resting or tasking. Six disease-related datasets contain about 1,500 subjects under the same resting state but various health status due to different brain-environment interactions.

To comprehensively evaluate and showcase the performance of the proposed LCM, we conduct experiments on both randomly initialized and pretrained models across clinical applications, as well as tasks involving sex, and cognitive state recognition. The fewshot finetuning experiments are also conducted. Pretrained models are finetuned with different ratio of data in the same validation fold to demonstrate performance for real-world clinical applications.

Datasets

We partition brain regions using the AAL atlas (Tzourio-Mazoyer et al. 2002) through all experiments. The data preprocessing details can refer to the extended version and a benchmark paper (Xu et al. 2023).

The Lifespan Human Connectome Project Aging (HCPA) dataset (Bookheimer et al. 2019) is instrumental in task recognition research, offering a comprehensive view of the aging process. It includes data from 717 subjects, encompassing fMRI records (n=4,863) with human behaviors associated with memory, sensory-motor and the resting state. In our experiments, these tasks are treated as a four-class classification.

The Human Connectome Project Young Adult (HCPYA) dataset (Van Essen et al. 2013) has tackled key aspects of the neural pathways that underlie brain function and behavior via high-quality neuroimaging data in over 1100 healthy young adults. It includes data from seven human behaviors associated with various cognitive tasks, e.g., language and working memory. In our experiments, these tasks are treated as a seven-class classification.

Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset (Weiner et al. 2015) serves as an invaluable resource, featuring a collection of pre-processed fMRI (n=138) and including clinical diagnostic labels. It encompasses a spectrum of cognitive states: Cognitive Normal (CN), Subjective Memory Complaints (SMC), Early-Stage Mild Cognitive Impairment (EMCI), Late-Stage Mild Cognitive Impairment (LMCI), and Alzheimer's Disease (AD). Considering the class unbalance issue, we simplified these categories into two broad groups based on disease severity: we combined CN, SMC, and EMCI into 'CN' group, while LMCI and AD were grouped as the 'AD' group.

Parkinson's Progression Markers Initiative (PPMI) dataset (Xu et al. 2023) presents a substantial collection of data from 209 subjects. It encompasses states of mental health: normal control, scans without evidence of dopamin-

	Pretrained on				Alzheimer's		Parkinson's		Autism	
	①	②	③	④	Accuracy	F1 score	Accuracy	F1 score	Accuracy	F1 score
BrainGNN					$83.48_{\pm 6.99}$	$78.69_{\pm 8.25}$	$76.92_{\pm 15.60}$	$75.24_{\pm 16.43}$	$62.83_{\pm 2.77}$	$61.73_{\pm 3.83}$
BolT					$82.89_{\pm 8.66}$	$77.66_{\pm 10.03}$	$73.30_{\pm 20.23}$	$69.93_{\pm 23.98}$	$68.09_{\pm 3.78}$	$68.01_{\pm 3.72}$
BNT					$83.56_{\pm 7.47}$	$78.31_{\pm 12.17}$	$76.79_{\pm 14.59}$	$71.99_{\pm 17.45}$	$70.23_{\pm 3.69}$	$69.98_{\pm 3.73}$
Graphormer					$82.81_{\pm 7.90}$	$77.51_{\pm 11.46}$	$65.89_{\pm 14.50}$	$61.84_{\pm 17.07}$	$57.66_{\pm 2.79}$	$57.00_{\pm 2.92}$
NAGphormer					$80.59_{\pm 6.82}$	$76.07_{\pm 9.72}$	$72.16_{\pm 18.01}$	$67.84_{\pm 19.93}$	$64.78_{\pm 2.47}$	$64.60_{\pm 2.73}$
NeuroPath					$82.07_{\pm 6.86}$	$74.12_{\pm 9.56}$	$69.79_{\pm 17.43}$	$65.81_{\pm 19.77}$	$65.71_{\pm 5.93}$	$64.64_{\pm 7.36}$
LCM†					$85.04_{\pm 9.30}$	$79.58_{\pm 13.77}$	$70.25_{\pm 14.36}$	$66.10_{\pm 16.62}$	$69.65_{\pm 5.10}$	$69.58_{\pm 5.23}$
BrainLM‡	✓				$83.30_{\pm 4.71}$	$75.79_{\pm 6.63}$	$50.43_{\pm 19.59}$	$45.83_{\pm 23.41}$	$53.25_{\pm 4.00}$	$51.51_{\pm 5.67}$
BrainLM	✓	✓	✓	✓	$82.56_{\pm 4.01}$	$75.41_{\pm 6.21}$	$54.40_{\pm 12.37}$	$50.37_{\pm 13.25}$	$49.90_{\pm 4.86}$	$33.89_{\pm 5.18}$
BrainMass-SVM	✓	✓	✓	✓	$82.96_{\pm 5.02}$	$75.32_{\pm 7.06}$	$59.21_{\pm 21.68}$	$51.25_{\pm 25.07}$	$64.49_{\pm 1.59}$	$64.56_{\pm 1.62}$
BrainMass-MLP	✓	✓	✓	✓	$82.96_{\pm 5.02}$	$75.32_{\pm 7.06}$	$73.78_{\pm 17.95}$	$70.20_{\pm 21.02}$	$68.48_{\pm 4.50}$	$68.30_{\pm 4.70}$
Brain-JEPA	✓	✓	✓	✓	$\underline{86.02_{\pm 3.49}}$	$80.20_{\pm 4.95}$	$\underline{82.09_{\pm 9.58}}$	$77.16_{\pm 12.34}$	$63.84_{\pm 0.82}$	$61.53_{\pm 1.35}$
	✓				$83.56_{\pm 7.92}$	$80.53_{\pm 9.57}$	$\mathbf{82.25}_{\pm 14.96}$	$\underline{82.97_{\pm 15.21}}$	$69.71_{\pm 0.43}$	$69.68_{\pm 0.54}$
	✓		✓	✓	$84.15_{\pm 3.50}$	$81.00_{\pm 5.59}$	$73.98_{\pm 14.69}$	$78.74_{\pm 12.19}$	$70.29_{\pm 3.26}$	$\underline{71.22_{\pm 2.85}}$
	✓	✓		✓	$84.30_{\pm 7.40}$	$82.52_{\pm 8.67}$	$79.00_{\pm 14.58}$	$82.68_{\pm 11.30}$	$69.51_{\pm 3.19}$	$70.48_{\pm 2.03}$
	✓	✓	✓		$84.89_{\pm 5.53}$	$\underline{83.48_{\pm 7.03}}$	$74.63_{\pm 16.48}$	$77.17_{\pm 14.66}$	$\underline{70.49_{\pm 1.76}}$	$70.82_{\pm 1.90}$
LCM	✓	✓	✓	✓	$\mathbf{86.30}_{\pm 5.80}$	$\mathbf{85.33}_{\pm 7.35}$	$81.30_{\pm 14.55}$	$\mathbf{84.18}_{\pm 11.63}$	$\mathbf{71.46}_{\pm 2.62}$	$\mathbf{72.50}_{\pm 1.91}$

Table 1: Finetune LCM with weights learned from various combination of BEIs. Diverse BEIs contribute differently to LCM pretraining, where checkmarks indicate which BEI (①: cognitive state, ②: Alzheimer's, ③: Parkinson's, and ④: Autism) are involved, † denotes LCM is not pretrained, ‡ denotes BrainLM is finetuned on the released model weights, **Bold** indicates the first ranking place, and underline indicates the second.

Sex↑	BrainLM	BM-SVM	BM-MLP	Brain-JEPA	LCM
HCPA	$40.68_{\pm 4.03}$	$69.44_{\pm 1.69}$	$\underline{69.93_{\pm 1.64}}$	$43.53_{\pm 0.65}$	$\mathbf{73.94}_{\pm 2.45}$
HCPYA	$38.97_{\pm 5.40}$	$68.12_{\pm 3.56}$	$\underline{68.54_{\pm 3.36}}$	$43.63_{\pm 3.27}$	$\mathbf{72.23}_{\pm 1.92}$
ADNI	$41.72_{\pm 7.19}$	$45.40_{\pm 12.88}$	$\underline{65.74_{\pm 7.66}}$	$62.34_{\pm 6.51}$	$\mathbf{71.98}_{\pm 6.87}$
ABIDE	$\underline{78.59_{\pm 4.98}}$	$73.84_{\pm 3.49}$	$74.73_{\pm 4.20}$	$78.11_{\pm 6.41}$	$\mathbf{87.34}_{\pm 4.48}$
PPMI	$42.03_{\pm 9.21}$	$48.83_{\pm 6.27}$	$\underline{67.73_{\pm 12.68}}$	$48.46_{\pm 7.09}$	$\mathbf{77.97}_{\pm 3.76}$
Taowu	$62.00_{\pm 23.53}$	$46.24_{\pm 23.96}$	$69.29_{\pm 17.74}$	$\mathbf{92.48}_{\pm 11.16}$	$\underline{90.67_{\pm 11.43}}$
Neurocon	$49.33_{\pm 28.08}$	$33.24_{\pm 15.34}$	$61.43_{\pm 27.90}$	$\underline{83.33_{\pm 12.43}}$	$\mathbf{100.00}_{\pm 0.00}$

Table 2: Performance on sex prediction across 7 datasets, where **Bold** indicates the first ranking place, and underline indicates the second, where BM refers BrainMass.

ergic deficit (SWEDD), prodromal, and Parkinson's disease (PD). In our experiments, the dataset is treated as a four-class classification.

Autism Brain Imaging Data Exchange (ABIDE) dataset presents data from 1025 young adults. The initiative aggregated fMRI data collected from laboratories around the world to support the research on Autism Spectrum Disorder (ASD). Subjects are classified into typical controls and those suffering from ASD. The binary classification is set for this dataset in our experiments.

Taowu and **Neurocon** (Xu et al. 2023) are two of the earliest image datasets released for Parkinson's and contain 81 subjects. The binary classification is set for these datasets in our experiments.

Schizophrenia (SZ) contains 189 subjects. There are 30 diseased and 159 healthy. The binary classification is set for the dataset.

Implementation Details

Following previous works, our experiments are done with subject-level cross-validation (CV). The average score and the standard deviation are both listed. To make our results comparable with previous papers, HCPA, HCPYA, and ADNI use a 5-fold CV as same as (Dan et al. 2023; Wei et al. 2024), while others use 10-fold as same as (Xu et al. 2023). Since LCM is a foundation model, training data for pretraining and finetuning is always from the corresponding CV fold's training set to prevent data leakage. Hyperparameters, e.g., learning rate and hidden channels, can be found in the extended version. SOTA brain-dedicated models, BrainGNN (Li et al. 2021), BNT (Kan et al. 2022), BolT (Bedel et al. 2023), and NeuroPath (Wei et al. 2024) are implemented as their original codes with default hyperparameters, where the structural connectome utilized by NeuroPath is replaced by FC in our work. Additionally, SOTA graph Transformers, Graphormer (Ying et al. 2021), and NAGphormer (Chen et al. 2023), are also compared. The released BrainLM (Ortega Caro et al. 2023), the one we trained from scratch, along with the original BrainMass-SVM (Yang et al. 2024), the modified BrainMass-MLP, and Brain-JEPA (Dong et al. 2024), are both retrained and compared. Codes and model weights can be found in the extended version.

Main Results: Finetuning Performance

Disease diagnosis As listed in Table 1, there are four additional versions of pretraining LCM, (i): without any diseases, (ii): no Alzheimer's, (iii): no Parkinson's, and (iv) no Autism, to compare with LCM pretraining with all data in the last row. SOTA models are also listed in the upper part of the Table for comparison. Note that, due to class unbalance, accuracy can have different ranks as F1 score, and F1 is the metric to conclude.

Firstly, pretraining with more data leads to better performance, and hence, the complete version of LCM has the best accuracy and F1 score against others. LCM pretrained with one or more disease-related datasets can outperform the one pretrained with only HCPs (①) except for Parkinson's,

	HCPA 3-task	HCPYA 7-task	ADNI Alzheimer's	PPMI Parkinson's	ABIDE Autism	Taowu Parkinson's	Neurocon Parkinson's
MLP-Small	93.99$_{\pm 0.35}$	88.67$_{\pm 2.07}$	76.39$_{\pm 8.90}$	61.45$_{\pm 11.05}$	68.81$_{\pm 3.04}$	61.00$_{\pm 28.29}$	65.87$_{\pm 31.40}$
MLP-Mid	91.87$_{\pm 1.09}$	80.04$_{\pm 3.40}$	75.85$_{\pm 8.07}$	58.32$_{\pm 15.46}$	66.14$_{\pm 6.56}$	50.67$_{\pm 29.09}$	62.48$_{\pm 29.39}$
MLP-Big	87.37$_{\pm 8.25}$	76.43$_{\pm 1.86}$	76.17$_{\pm 7.36}$	41.74$_{\pm 12.82}$	49.21$_{\pm 14.90}$	43.00$_{\pm 31.84}$	59.54$_{\pm 30.60}$
LCM-Small	97.04$_{\pm 0.40}$	94.63$_{\pm 0.77}$	76.99$_{\pm 9.91}$	57.29$_{\pm 14.52}$	68.34$_{\pm 3.54}$	74.57$_{\pm 27.90}$	59.95$_{\pm 27.60}$
LCM-Mid	97.15$_{\pm 0.25}$	94.84$_{\pm 0.61}$	77.96$_{\pm 13.41}$	58.19$_{\pm 13.35}$	69.39$_{\pm 5.02}$	78.33$_{\pm 20.39}$	69.21$_{\pm 28.19}$
LCM-Big	**97.18**$_{\pm 0.54}$	**95.02**$_{\pm 1.00}$	**79.58**$_{\pm 13.77}$	**61.61**$_{\pm 14.69}$	**69.58**$_{\pm 5.23}$	**88.33**$_{\pm 22.52}$	**71.87**$_{\pm 27.42}$

Table 3: Model scalability demonstration via performance on phenotypic prediction. **Bold** indicates the first ranking place, and underline indicates the second.

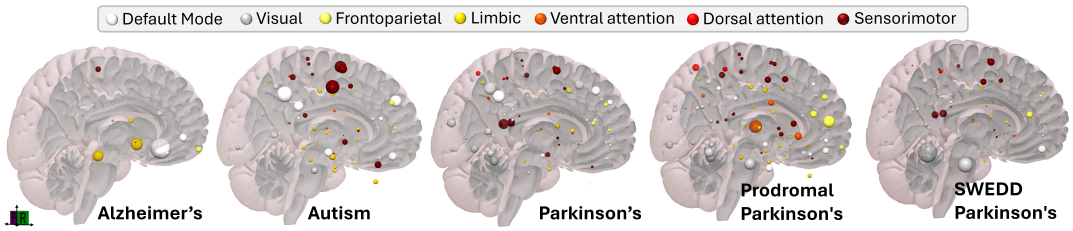

Figure 6: The average cross-attention map of all test data at the readout layer of LCM on disease-related datasets. The node size indicates the relative attention weight.

Schizophrenia (SZ)	10% FT	50% FT	100% FT
BolT	79.70$_{\pm 9.76}$	80.16$_{\pm 12.68}$	81.14$_{\pm 9.75}$
NeuroPath	78.60$_{\pm 9.77}$	80.16$_{\pm 12.68}$	81.74$_{\pm 8.90}$
BrainMass-SVM	78.60$_{\pm 8.74}$	78.60$_{\pm 8.74}$	78.60$_{\pm 8.74}$
BrainMass-MLP	78.89$_{\pm 10.14}$	80.95$_{\pm 11.00}$	82.55$_{\pm 10.95}$
LCM	**81.63**$_{\pm 8.58}$	**81.73**$_{\pm 11.48}$	**83.61**$_{\pm 6.01}$
p-value	0.0267	0.0124	0.0014

Table 4: Fewshot finetuning (FT) performance comparison on a held-out disease.

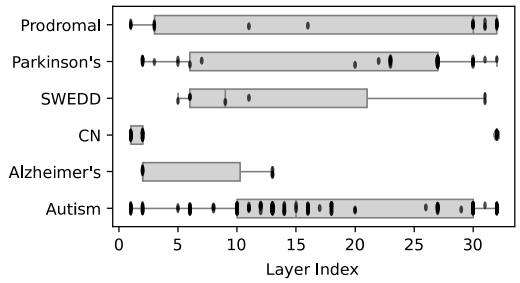

Figure 7: The distribution of the index of the best-matched readout layer of LCM, where a dot in the box plot represents one sample point.

which is the only four-class classification task and has higher difficulty.

Additionally, pretraining enhances the LCM performance even if the downstream tasks are unseen. After holding out all disease-related datasets (pretrained on ①), the F1 scores on all diseases are all better than the train-from-scratch LCM†. Performance improvement is impressive in Parkinson's, given a 16.87% increase in F1, while before pretraining LCM† is about 11% lower than Brain-JEPA in F1. Compared to LCM†, LCM is always better in F1. We can find this observation for cases of held-out Alzheimer's (pretrained on ① ③ ④), Parkinson's (pretrained on ① ② ④), or Autism (pretrained on ① ② ③).

Last but not least, finetuning the LCM on seen datasets is more stable than on unseen datasets. Take Alzheimer's as an example, pretrained on ① ② ③ and ① ② ④ have 82.52% and 83.48% F1, respectively, while the best of unseen is 81%. Autism also shows better when pretrained on ① ③ ④ than on ① ② ③ in F1 scores. Whilst Parkinson's shows the opposite due to the difficult four-class classification.

Demographics We test models to predict the sex on our seven datasets, respectively, as listed in Table 2, where sex is measured by F1 score. Clearly, LCM outperforms others on all datasets except for Taowu ($n = 40$), while other models are unable to hold the first/second place across all datasets.

Main Results: Fewshot Finetuning

To further demonstrate the generalizability, we finetune and evaluate LCM on a new unseen dataset, SZ ($n = 189$) by F1 score as listed in Table 4, p-value is from a paired t-test between scores of LCM and other SOTA models in the 5-fold cross-validation. As the experiments in BrainMass, different ratios of finetuning data, 10%, 50%, and 100%, were used in the evaluation. We can see the generalizability of LCM is significantly better ($p < 0.05$) than BrainMass.

Main Results: Model Scalability

The comparison of classification performance measured by F1 score is listed in Table 3, where Small refers to 8-layer (100M-level), Mid is 20-layer (700M-level), and Big is 32-layer (1.2B-level). In contrast, SOTA models in the middle

	Supervised at			Alzheimer's		Parkinson's		Autism	
	Last	Best	All	Accuracy	F1 score	Accuracy	F1 score	Accuracy	F1 score
Baseline	✓			$83.22_{\pm 5.72}$	$80.49_{\pm 6.69}$	$76.06_{\pm 16.02}$	$78.45_{\pm 14.31}$	$\underline{69.35_{\pm 3.22}}$	$\underline{69.90_{\pm 3.34}}$
Only stage 1			✓	$79.19_{\pm 6.33}$	$80.49_{\pm 6.63}$	$\underline{80.40_{\pm 11.13}}$	$\mathbf{84.74}_{\pm 8.35}$	$65.24_{\pm 2.86}$	$66.41_{\pm 2.25}$
Only stage 2		✓		$\underline{84.15_{\pm 3.50}}$	$\underline{83.78_{\pm 6.69}}$	$75.35_{\pm 13.80}$	$80.95_{\pm 7.68}$	$66.80_{\pm 2.10}$	$67.32_{\pm 1.76}$
Stage 1 & 2		✓	✓	$\mathbf{86.30}_{\pm 5.80}$	$\mathbf{85.33}_{\pm 7.35}$	$\mathbf{81.30}_{\pm 14.55}$	$\underline{84.18_{\pm 11.63}}$	$\mathbf{71.46}_{\pm 2.62}$	$\mathbf{72.50}_{\pm 1.91}$

Table 5: Ablation studies of LCM learning strategy, where average scores are listed.

rows have only 4M-level parameters. Note that although either width or depth can increase the model scale of MLP, we found that width is not as valuable as the depth of MLP, referring to the similar accuracy with different widths as shown in the extended version. Except for BrainLM and LCM-Big which are single model pretrained with additional data and tested on each dataset, all models listed in Table 3 are trained from scratch for each dataset.

Clearly, LCM holds the best/second performance given the best scalability, where the F1 score is always increased when the model size is enlarged, and it is finally significantly boosted by the pretraining. MLP is not scalable, agreeing with the exploded training loss as shown in Fig. 2. The best performance by MLP is consistently demonstrated by the small version, while larger MLPs have lower F1 scores. For example, MLP-Small ranks in the top 5 for PPMI and ABIDE, but MLP-Big drops to last place. More analysis of MLP scalability can be found in extended version.

It is worth noting that LCM-Big got enhanced to the best performance with over 90% F1 on Taowu and Neurocon datasets by pretraining (see Table 1), while before that, LCM-Big was out of the top three. LCM drops on HCPA and HCPYA after pretraining because their sample size is 10 times the rest of the datasets. However, the disease-related datasets that have small sample sizes consistently benefited from big data.

Interpretation

Adaptive training We demonstrate the frequency of the best-matched readout layer index across batches by box plot in Fig 7. We can see it has a diverse distribution for various disease diagnoses. Alzheimer's, Control Normal (CN), and SWEDD tend to use shallow features at layers between 1 to 15 across batches since it is relatively easier to separate Alzheimer's (dementia stage) and CN/SWEDD (with non-evident symptoms). In contrast, multi-level feature representations are required from various layers to effectively differentiate Autism, Parkinson's, and Prodromal stages because of subtle variations in brain function. This distribution is collected from the last epoch of finetuning, indicating semi-supervised LCM did not converged to a fixed layer for different phenotypes.

Visualizations The average cross-attention map of test data with the same label in ADNI, ABIDE, and PPMI datasets is shown in Fig. 6. The attention weights are extracted at the readout layer by the proposed adaptive training. We can observe that LCM is more attentive to the default mode network for Alzheimer's and Autism than Parkinson's, which aligns with current neuroscience knowledge (Padmanabhan et al. 2017; Zhang et al. 2023). It is clear that LCM is also attentive to the limbic network for ADNI and ABIDE, which is often damaged in Alzheimer's disease (Hopper and Vogel 1976) and Autism (Wong et al. 2020). For three classes of Parkinson's, LCM is attentive to sensorimotor, visual, and frontoparietal networks that are involved in Parkinson's disease by agreeing with (Schneider, Diamond, and Markham 1987; Cascone et al. 2021; Göttlich et al. 2013). These visualizations can interpret the promising performance of LCM.

Ablation studies

As introduced in Section 3, the learning strategy of LCM is designed with a momentum of full supervision (stage 1) on all decoder layers in a few beginning epochs, and then followed by an adaptive training (stage 2) on the best-matched layer. To show the effectiveness of the learning strategy design, we run finetuning experiments for the pretrained LCM with all three ablate version as listed in Table 5, where the model supervised at the last layer as the baseline represents the training of a generic Transformer decoder. Only stage 2 across the entire finetuning shows the necessity of the momentum of full supervision at stage 1, and only stage 1 shows the necessity of stage 2.

Clearly, as listed in Table 5, our design of training LCM shows the best performance on disease-related datasets. Other versions of LCM finetuning are not as stable as the performance gained by our design. Although the only stage 2 version shows the second-best performance on Alzheimer's, it has worse scores than the train-from-scratch LCM on Autism (see Table 1). As well as the only stage 1 version acts well on Parkinson's but fails on Autism. This observation supports that different phenotypes perform great at different layers of LCM, and momentum by full supervision is required for training LCM on the correct direction.

5 Conclusion

In conclusion, we proposed a large connectome model (LCM), which is the largest brain foundation model (1.2B) for clinical applications. We have evaluated our foundation model on a variety of applications, including sex prediction, human behavior recognition, and early diagnosis of Autism, Parkinson's disease, and Alzheimer's disease, where promising results shown by the pretrained LCM indicate the great potential to facilitate brain connectome in clinical routines. The LCM finetuning on unseen datasets is also promoted by the pretraining, with significant performance enhancement compared to train-from-scratch LCM. Given the impressive performance of our methods on 8 datasets, the decoder-only architecture learning from the multitask learning provides a new routine for training a brain foundation model.

Acknowledgements

This work was supported by the National Institutes of Health (AG091653, AG068399, AG084375) and the Foundation of Hope.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Assran, M.; Duval, Q.; Misra, I.; Bojanowski, P.; Vincent, P.; Rabbat, M.; LeCun, Y.; and Ballas, N. 2023. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15619–15629.

Bedel, H. A.; Sivgin, I.; Dalmaz, O.; Dar, S. U.; and Çukur, T. 2023. BolT: Fused window transformers for fMRI time series analysis. *Medical Image Analysis*, 88: 102841.

Bookheimer, S. Y.; Salat, D. H.; Terpstra, M.; Ances, B. M.; Barch, D. M.; Buckner, R. L.; Burgess, G. C.; Curtiss, S. W.; Diaz-Santos, M.; Elam, J. S.; et al. 2019. The lifespan human connectome project in aging: an overview. *Neuroimage*, 185: 335–348.

Caballero-Gaudes, C.; and Reynolds, R. C. 2017. Methods for cleaning the BOLD fMRI signal. *Neuroimage*, 154: 128–149.

Cascone, A. D.; Langella, S.; Sklerov, M.; and Dayan, E. 2021. Frontoparietal network resilience is associated with protection against cognitive decline in Parkinson's disease. *Communications biology*, 4(1): 1021.

Chen, J.; Gao, K.; Li, G.; and He, K. 2023. NAGphormer: A Tokenized Graph Transformer for Node Classification in Large Graphs. In *Proceedings of the International Conference on Learning Representations*.

Cui, H.; Dai, W.; Zhu, Y.; Kan, X.; Gu, A. A. C.; Lukemire, J.; Zhan, L.; He, L.; Guo, Y.; and Yang, C. 2022. Braingb: a benchmark for brain network analysis with graph neural networks. *IEEE transactions on medical imaging*, 42(2): 493–506.

Dan, T.; Ding, J.; Wei, Z.; Kovalsky, S.; Kim, M.; Kim, W. H.; and Wu, G. 2023. Re-Think and Re-Design Graph Neural Networks in Spaces of Continuous Graph Diffusion Functionals. *Advances in Neural Information Processing Systems*, 36: 59375–59387.

Ding, J.; Dan, T.; Wei, Z.; Cho, H.; Laurienti, P. J.; Kim, W. H.; and Wu, G. 2024. Machine Learning on Dynamic Functional Connectivity: Promise, Pitfalls, and Interpretations. *arXiv preprint arXiv:2409.11377*.

Dong, Z.; Li, R.; Wu, Y.; Nguyen, T. T.; Chong, J. S. X.; Ji, F.; Tong, N. R. J.; Chen, C. L. H.; and Zhou, J. H. 2024. Brain-JEPA: Brain Dynamics Foundation Model with Gradient Positioning and Spatiotemporal Masking. *arXiv preprint arXiv:2409.19407*.

Göttlich, M.; Münte, T. F.; Heldmann, M.; Kasten, M.; Hagenah, J.; and Krämer, U. M. 2013. Altered resting state brain networks in Parkinson's disease. *PloS one*, 8(10): e77336.

He, T.; An, L.; Chen, P.; Chen, J.; Feng, J.; Bzdok, D.; Holmes, A. J.; Eickhoff, S. B.; and Yeo, B. T. 2022. Meta-matching as a simple framework to translate phenotypic predictive models from big to small data. *Nature neuroscience*, 25(6): 795–804.

Hopper, M.; and Vogel, F. 1976. The limbic system in Alzheimer's disease. A neuropathologic investigation. *The American journal of pathology*, 85(1): 1.

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

Keriven, N. 2022. Not too little, not too much: a theoretical analysis of graph (over) smoothing. *Advances in Neural Information Processing Systems*, 35: 2268–2281.

Li, X.; Zhou, Y.; Dvornek, N.; Zhang, M.; Gao, S.; Zhuang, J.; Scheinost, D.; Staib, L. H.; Ventola, P.; and Duncan, J. S. 2021. Braingnn: Interpretable brain graph neural network for fmri analysis. *Medical Image Analysis*, 74: 102233.

Ortega Caro, J.; Oliveira Fonseca, A. H.; Averill, C.; Rizvi, S. A.; Rosati, M.; Cross, J. L.; Mittal, P.; Zappala, E.; Levine, D.; Dhodapkar, R. M.; et al. 2023. BrainLM: A foundation model for brain activity recordings. *bioRxiv*, 2023–09.

Padmanabhan, A.; Lynch, C. J.; Schaer, M.; and Menon, V. 2017. The default mode network in autism. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 2(6): 476–486.

Paul, D.; Chowdhury, A.; Xiong, X.; Chang, F.-J.; Carlyn, D. E.; Stevens, S.; Provost, K. L.; Karpatne, A.; Carstens, B.; Rubenstein, D.; Stewart, C.; Berger-Wolf, T.; Su, Y.; and Chao, W.-L. 2024. A Simple Interpretable Transformer for Fine-Grained Image Classification and Analysis. In *The Twelfth International Conference on Learning Representations*.

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

Said, A.; Bayrak, R.; Derr, T.; Shabbir, M.; Moyer, D.; Chang, C.; and Koutsoukos, X. 2023. Neurograph: Benchmarks for graph machine learning in brain connectomics. *Advances in Neural Information Processing Systems*, 36: 6509–6531.

Schneider, J. S.; Diamond, S. G.; and Markham, C. H. 1987. Parkinson's disease: sensory and motor problems in arms and hands. *Neurology*, 37(6): 951–951.

Tzourio-Mazoyer, N.; Landeau, B.; Papathanassiou, D.; Crivello, F.; Etard, O.; Delcroix, N.; Mazoyer, B.; and Joliot, M. 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage*, 15(1): 273–289.

Van Essen, D. C.; Smith, S. M.; Barch, D. M.; Behrens, T. E.; Yacoub, E.; Ugurbil, K.; Consortium, W.-M. H.; et al. 2013. The WU-Minn human connectome project: an overview. *Neuroimage*, 80: 62–79.

Wei, Z.; Dan, T.; Ding, J.; and Wu, G. 2024. NeuroPath: A Neural Pathway Transformer for Joining the Dots of Human Connectomes. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Weiner, M. W.; Veitch, D. P.; Aisen, P. S.; Beckett, L. A.; Cairns, N. J.; Cedarbaum, J.; Donohue, M. C.; Green, R. C.; Harvey, D.; Jack Jr, C. R.; et al. 2015. Impact of the Alzheimer's disease neuroimaging initiative, 2004 to 2014. *Alzheimer's & Dementia*, 11(7): 865–884.

Wen, G.; Cao, P.; Liu, L.; Yang, J.; Zhang, X.; Wang, F.; and Zaiane, O. R. 2023. Graph self-supervised learning with application to brain networks analysis. *IEEE Journal of Biomedical and Health Informatics*, 27(8): 4154–4165.

Wong, N. M.; Findon, J. L.; Wichers, R. H.; Giampietro, V.; Stoencheva, V.; Murphy, C. M.; Blainey, S.; Ecker, C.; Murphy, D. G.; McAlonan, G. M.; et al. 2020. Serotonin differentially modulates the temporal dynamics of the limbic response to facial emotions in male adults with and without autism spectrum disorder (ASD): a randomised placebo-controlled single-dose crossover trial. *Neuropsychopharmacology*, 45(13): 2248–2256.

Xu, J.; Yang, Y.; Huang, D.; Gururajapathy, S. S.; Ke, Y.; Qiao, M.; Wang, A.; Kumar, H.; McGeown, J.; and Kwon, E. 2023. Data-driven network neuroscience: On data collection and benchmark. *Advances in Neural Information Processing Systems*, 36: 21841–21856.

Yang, Y.; Ye, C.; Su, G.; Zhang, Z.; Chang, Z.; Chen, H.; Chan, P.; Yu, Y.; and Ma, T. 2024. BrainMass: Advancing Brain Network Analysis for Diagnosis with Large-scale Self-Supervised Learning. *IEEE Transactions on Medical Imaging*.

Ying, C.; Cai, T.; Luo, S.; Zheng, S.; Ke, G.; He, D.; Shen, Y.; and Liu, T.-Y. 2021. Do transformers really perform badly for graph representation? *Advances in neural information processing systems*, 34: 28877–28888.

Zhang, Z.; Chan, M. Y.; Han, L.; Carreno, C. A.; Winter-Nelson, E.; Wig, G. S.; (ADNI, A. D. N. I.; et al. 2023. Dissociable effects of Alzheimer's disease-related cognitive dysfunction and aging on functional brain network segregation. *Journal of Neuroscience*, 43(46): 7879–7892.