PTGB: Pre-Train Graph Neural Networks for Brain Network Analysis

Yi Yang YI.YANG@EMORY.EDU
Emory University, USA

Hejie Cui HEJIE.CUI@EMORY.EDU
Emory Unversity, USA

Carl Yang J.CARLYANG@EMORY.EDU
Emory Unversity, USA

Abstract

The human brain is the central hub of the neurobiological system, controlling behavior and cognition in complex ways. Recent advances in neuroscience and neuroimaging analysis have shown a growing interest in the interactions between brain regions of interest (ROIs) and their impact on neural development and disorder diagnosis. As a powerful deep model for analyzing graph-structured data, Graph Neural Networks (GNNs) have been applied for brain network analysis. However, training deep models requires large amounts of labeled data, which is often scarce in brain network datasets due to the complexities of data acquisition and sharing restrictions. To make the most out of available training data, we propose PTGB, a GNN pre-training framework that captures intrinsic brain network structures, regardless of clinical outcomes, and is easily adaptable to various downstream tasks. PTGB comprises two key components: (1) an unsupervised pre-training technique designed specifically for brain networks, which enables learning from large-scale datasets without task-specific labels; (2) a data-driven parcellation atlas mapping pipeline that facilitates knowledge transfer across datasets with different ROI systems. Extensive evaluations using various GNN models have demonstrated the robust and superior performance of PTGB compared to baseline methods.

Data and Code Availability The empirical study in this work uses three real-world brain network datasets: 1) the Bipolar Disorder (BP) dataset, 2) the Human Immunodeficiency Virus Infection (HIV) dataset, and 3) the Parkinson's Progression Markers Initiative (PPMI) dataset. The BP and HIV are local datasets, while the large-scale PPMI dataset[1] is publicly available for authorized users. We followed the data preprocessing pipelines provided by the open-source BrainGB platform[2] (Cui et al., 2022a) for the construction of brain networks based on raw neuroimaging data. The full implementation of this work is publicly available at https://github.com/Owen-Yang-18/BrainNN-PreTrain.

Institutional Review Board (IRB) The study has been approved by an Institutional Review Board (IRB) to ensure the ethical and responsible use of human subjects in research. The IRB reviewed and approved the study protocols and consent forms, ensuring that the rights and welfare of the participants are protected. The study strictly adheres to the Good Clinical Practice guidelines and U.S. 21 CFR Part 50 (Protection of Human Subjects) to ensure the safety and privacy of the participants. All the data used in this work is processed anonymously to protect the privacy of participants, and no personally identifiable information is used or disclosed.

1. Introduction

Brain network analysis has attracted considerable interest in neuroscience studies in recent years. A brain network is essentially a connected graph constructed from different raw imaging modalities such as Diffusion Tensor Imaging (DTI) and functional Magnetic Resonance Imaging (fMRI), where nodes are composed by the anatomical regions of interest (ROIs) given predefined parcellation atlas, and connections are usually formed with the correlations among ROIs.

1. https://www.ppmi-info.org/
2. https://braingb.us/

Effective brain network analysis plays a pivotal role in understanding the biological structures and functions of complex neural systems, which potentially helps the early diagnosis of neurological disorders and facilitates neuroscience research (Martensson et al., 2018; Yahata et al., 2016; Lindquist, 2008; Smith, 2012).

Graph Neural Networks (GNNs) have emerged as a powerful tool for analyzing graph-structured data, delivering impressive results on a wide range of network datasets, including social networks, recommender systems, knowledge graphs, protein and gene networks, and molecules, among others (Kipf and Welling, 2017; Hamilton et al., 2017; Schlichtkrull et al., 2018; Vashishth et al., 2020; Xu et al., 2019; Ying et al., 2018; Zhang et al., 2020; Liu et al., 2022; Xiong et al., 2020; Cui et al., 2022d; Xu et al., 2022). These models have proven their ability to learn powerful representations and efficiently compute complex graph structures, making them well-suited for various downstream tasks. In the field of neuroscience, GNN has been applied to brain network analysis, specifically for graph-level classification/regression (Ying et al., 2018; Xu et al., 2019; Errica et al., 2020; Luo et al., 2022; Dai et al., 2023; Xu et al., 2023a) and important vertex/edge identification (Ying et al., 2019; Luo et al., 2020; Vu and Thai, 2020; Yu et al., 2023; Kan et al., 2022c), towards tasks such as connectome-based disease prediction and multi-level neural pattern discovery. However, deep learning models, including GNNs, require large amounts of labeled data to achieve optimal performance (Hu et al., 2020a; You et al., 2020; Zhu et al., 2021a). While neuroimaging datasets are available from national neuroimaging studies such as the ABCD (Casey et al., 2018), ADNI (Hinrichs et al., 2009), and PPMI (Aleksovski et al., 2018), these datasets are still relatively small compared to graph datasets from other domains, such as datasets with 41K to 452K graphs on OGB (Hu et al., 2020b) and datasets with thousands to millions of graphs on NetRepo (Rossi and Ahmed, 2016)). The limited amount of data can result in overfitting when training deep models.

Transfer learning offers a solution to the challenge of limited data availability in training deep models. It allows a model pre-trained on large-scale source datasets to be adapted to smaller target datasets while maintaining robust performance. However, the success of transfer learning depends on the availability of similar supervision labels on the source and target dataset. This is not always feasible in large-scale public studies, particularly in the field of brain network analysis. Self-supervised pre-training has been shown to be effective in various domains, such as computer vision (He et al., 2020; Chen et al., 2020b), natural language processing (Devlin et al., 2019; Yu et al., 2022), and graph mining (Sun et al., 2022). We aim to explore a self-supervised pre-training approach for GNNs on brain networks that is not restricted by task-specific supervision labels. Despite the promising potential, unique challenges still need to be addressed to achieve effective disease prediction. One of the major challenges is the inconsistent ROI parcellation systems in constructing different brain network datasets, which hinders the transferability of pre-trained models across datasets. The process of parcellating raw imaging data into brain networks is highly complex and usually done ad hoc by domain experts for each study, making it unrealistic to expect every institution to follow the same parcellation system. Although some institutions may release preconstructed brain network datasets (Di Martino et al., 2014), the requirement for universal adherence to a single parcellation system is infeasible.

To tackle the challenge of insufficient training data for GNNs in brain network analysis, we present **P**re-**T**raining **G**raph neural networks for **B**rain networks (PTGB), a fully unsupervised pre-training approach that captures shared structures across brain network datasets. PTGB adapts the data-efficient MAML (Finn et al., 2017) with a two-level contrastive learning strategy based on the naturally aligned node systems of brain networks across individuals. Additionally, to overcome the issue of diverse parcellation systems, we introduce a novel data-driven atlas mapping technique. This technique transforms the original features into low-dimensional representations in a uniform embedding space and aligns them using variance-based projection, which incorporates regularizations that preserve spatial relationships, consider neural modules, and promote sparsity.

In summary, our contributions are three-folded:

- We present an unsupervised pre-training approach for GNNs on brain networks, addressing the issue of resource-limited training.
- We propose a two-level contrastive sampling strategy tailored for GNN pre-training on brain networks, which combines with a data-driven brain atlas mapping strategy that employs customized regularizations and variance-based sorting to enhance cross-dataset learning.
- Our experiments against shallow and deep baselines demonstrate the effectiveness of our proposed

PTGB. Further, we provide an in-depth analysis to understand the influence of each component.

2. Related Work

GNNs for Brain Network Analysis. GNNs are highly effective for analyzing graph-structured data and there have been some pioneering attempts to use them for predicting diseases by learning over brain networks. For example, BrainGNN (Li et al., 2021b) proposes ROI-aware graph convolutional layers and ROI-selection pooling layers for predicting neurological biomarkers. BrainNetCNN (Kawahara et al., 2017) designs a CNN that includes edge-to-edge, edge-to-node, and node-to-graph convolutional filters, leveraging the topological locality of brain connectome structures. BrainNetTF (Kan et al., 2022b) introduces a transformer architecture with an orthonormal clustering readout function that considers ROI similarity within functional modules. Additionally, various studies (Cui et al., 2022c; Kan et al., 2022a; Zhu et al., 2022a; Cui et al., 2022a; Yu et al., 2023) have shown that, when data is sufficient, GNNs can greatly improve performance in tasks such as disease prediction. However, in reality, the lack of training data is a common issue in neuroscience research, particularly for specific domains and clinical tasks (Xu et al., 2023b). Despite this, there has been little research into the ability of GNNs to effectively train for brain network analysis when data is limited.

Unsupervised Graph Representation Learning and GNN Pre-training. Unsupervised learning is a widely used technique for training complex models when resources are limited. Recent advancements in contrastive learning (Chen et al., 2020a; He et al., 2020; Yu et al., 2021; Zhu et al., 2022b) have led to various techniques for graphs. For instance, GBT (Bielak et al., 2022) designs a Barlow Twins Zbontar et al. (2021) loss function based on the empirical cross-correlation of node representations learned from two different views of the graph (Zhao et al., 2021). Similarly, GraphCL (You et al., 2020) involves a comparison of graph-level representations obtained from two different augmentations of the same graph. DGI (Velickovic et al., 2019) contrasts graph and node representations learned from the original graph and its corruption.

To obtain strong models for particular downstream tasks, unsupervised training techniques can be used to pre-train a model, which is then fined tuned on the downstream tasks to reduce the dependence on labeled training data. The approach has proven highly successful in computer vision (Cao et al., 2020; Grill et al., 2020), natural language processing (Devlin et al., 2019; Radford et al., 2018, 2021; Liang et al., 2020), and multi-modality (e.g. text-image pair) learning (Li et al., 2022; Yao et al., 2022). There are various strategies for pre-training GNNs as well. GPT-GNN (Hu et al., 2020c) proposes graph-oriented pretext tasks, such as masked attribute and edge reconstruction. L2P-GNN (Lu et al., 2021) introduces dual adaptation by simultaneously optimizing the encoder on a node-level link prediction objective and a graph-level self-supervision task similar to DGI. Others, such as GMPT (Hou et al., 2022) adopt an inter-graph message-passing approach to obtain context-aware node embedding and optimize the model concurrently under supervision and self-supervision. To the best of our knowledge, the effectiveness of both contrastive learning and pre-training has not been investigated in the context of the unique properties of brain networks.

3. Unsupervised Brain Network Pre-training

Problem Definition. The available training resource includes a collection of brain network datasets $\mathcal{S} = \{\mathcal{D}_1, \mathcal{D}_2, \cdots \mathcal{D}_s\}$, where each dataset contains a varying number of brain networks. We consider each brain network instance with M number of defined ROIs as an undirected weighted graph $\mathcal{G}$ with M nodes. $\mathcal{G}$ is represented by a node-set $\mathcal{V} = \{v_m\}_{m=1}^{M}$, an edge set $\mathcal{E} = \mathcal{V} \times \mathcal{V}$, and a weighted adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{M \times M}$. We define a θ parameterized GNN model $f(\cdot)$, and our goal is to propose a pre-training schema that can effectively learn an initialization θ_0 for $f(\cdot)$ on a set of source datasets $\mathcal{S}_{\text{source}} \subset \mathcal{S}$ via self-supervision and adapt $f_{\theta_0}(\cdot)$ to a local optimum θ^* on a target set $\mathcal{S}_{\text{target}} \in \mathcal{S}$.

3.1. GNN Pre-training for Brain Networks

The goal of pre-training a GNN model for brain networks is to learn an appropriate initialization that can easily be adapted to downstream task. Note that the concept of pre-training is distinct from transfer learning since the latter expects a similarity between the source and target data as well as their learning objectives (*e.g.,* loss functions), while this is often lacking in brain network analysis due to absence of

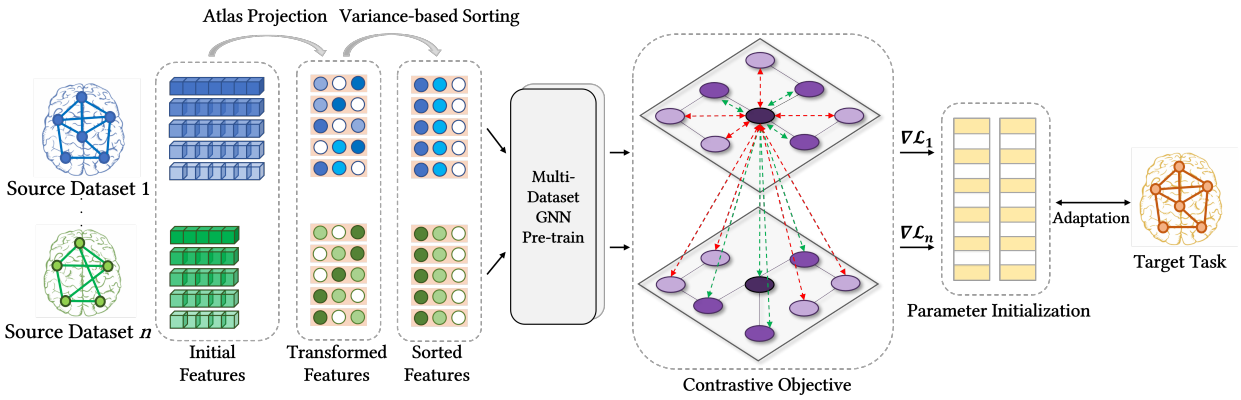


Figure 1: Overview of the proposed framework PTGB. The initial features of the source datasets are projected to a fixed dimension through atlas transformation followed by variance-based feature alignment, which facilitates self-supervised GNN pre-training on multiple datasets via the novel two-level contrastive learning objective. The learned model can serve as the parameter initialization and be further fine-tuned on target tasks.

sufficient ground truth labels in large scale studies as well as inherent differences in their brain network parcellation methods across varying datasets. Practically, a GNN model can be pre-trained either on a singular task with a single source dataset or on a collection of tasks with multiple source datasets. The proposed PTGB framework adopts the latter option since multi-task pre-training reduces the likelihood of the model being biased towards the knowledge of data from a singular source, which could be particularly concerning if the source and target data shares limited similarity leading to poor downstream adaptation due to information loss during model transfer. However, a naive approach towards multi-task pre-training would not suffice in learning a robust model initialization. Specifically, it presents two underlying risks: (1) the model may not perform consistently well on all tasks and may also overfit to a particular task which significantly undermines model generalizability; and (2) the process could be computationally inefficient with increasing number of tasks regardless if the model is optimized sequentially or simultaneously on all tasks (Yang et al., 2022).

To this end, we adopt the popular data-efficient training techniques presented in MAML (Finn et al., 2017) with the goal of ensuring consistent performance on all tasks as well as computational efficiency. The MAML technique is characterized by an inner-loop adaptation and an outer-loop update (Raghu et al., 2019). At each training iteration, each input dataset is partitioned into an inner-loop support set and an outer-loop query set. The model is first trained on the support set without explicitly updating the parameters. Instead, the updates are temporarily stored as fast weights (Ba et al., 2016). These fast weights are then used to evaluate the query set and compute the actual gradients. This approach makes use of approximating higher-order derivatives (Tan and Lim, 2019) at each step, allowing the model to foresee its optimization trajectory a few steps ahead, which practically reduces the number of required training iterations to reach local optima. In our scenario, the joint optimization involves summing the loss over each brain network dataset, i.e., for n number of datasets and their respective temporary fast weights $\{\theta_i'\}_{i=1}^n$ and outer-loop queries $\{\text{query}_i\}_{i=1}^n$, the step-wise update of the model parameter at time t is $\theta^{t+1} = \theta^t - \alpha\nabla_{\theta^t}\sum_{i=1}^n \mathcal{L}_{\text{query}_i} f_{\theta_i'^t}(\cdot)$. We hereby summarize this process in Algorithm 1. In addition, we will also demonstrate the advantages of MAML-styled pre-training over vanilla multi-task pre-training as well as single task pre-training through experiments which will be discussed in Section 4.1.

3.2. Brain Network Oriented Two-Level Contrastive Learning

Given the high cost of acquiring labeled training data for brain network analysis, our pre-training pipeline

Algorithm 1 MAML-based Multi-task Pre-training

Input: Source task pool S_τ, GNN model $f_\theta(\cdot)$
Require: α, β: learning rate hyperparameters
Randomly initialize θ
while *not done* **do**
  for *each task* τ_i *in* S_τ **do**
    Sample a set of k datapoints $\mathcal{D}_i$ from τ_i as support set
    Evaluate the gradient for the task-wise objective $\nabla_\theta \mathcal{L}_{\mathcal{D}_i} f(\theta)$
    Compute the inner-loop adapted parameters $\theta_i' \leftarrow \theta - \beta \nabla_\theta \mathcal{L}_{\mathcal{D}_i} f(\theta)$
    Sample another set of datapoints $\mathcal{D}_i'$ from τ_i as query set
  end
  Update the GNN model parameters $\theta \leftarrow \theta - \alpha \nabla_\theta \sum_{\mathcal{D}_i' \sim S_\tau} \mathcal{L}_{\mathcal{D}_i'} f_{\theta_i'}(\cdot)$
end

of PTGB adopts to the effective label-free learning strategy of contrastive learning (CL). CL aims to maximize the mutual information (MI) between an anchor point of investigation X from a data distribution $\mathcal{H}$ and its positive samples X^+, while minimizing MI with its negative samples X^-. The contrastive objective function is formulated as follows:

$$\mathcal{J}_{\text{con}} = \arg\min \left[\left(-I(X; X^+) + I(X; X^-) \right) \right]. \quad (1)$$

In the context of graph CL, given an anchor node representation z_α, a set of positive samples $\mathbf{S}^+$, and a set of negative samples $\mathbf{S}^-$, the training objective is based on the Jensen-Shannon divergence (Hjelm et al., 2019),

$$\mathcal{J}_{\text{JSD}}(z_\alpha) = \arg\min \left[\left(-I(z_\alpha; \mathbf{S}^+) + I(z_\alpha; \mathbf{S}^-) \right) \right], \quad (2)$$

where

$$I(z_\alpha; \mathbf{S}^+) = \frac{1}{|\mathbf{S}^+|} \sum_{z_{s^+} \in \mathbf{S}^+} \text{sp}\left(\frac{z_\alpha^\top z_{s^+}}{\|z_\alpha\| \|z_{s^+}\|} \right), \quad (3)$$

$$I(z_\alpha; \mathbf{S}^-) = \frac{1}{|\mathbf{S}^-|} \sum_{z_{s^-} \in \mathbf{S}^-} \text{sp}\left(\frac{z_\alpha^\top z_{s^-}}{\|z_\alpha\| \|z_{s^-}\|} \right), \quad (4)$$

and $\text{sp}(\cdot) = \log(1 + e^{\cdot})$ is softplus nonlinearity.

The ultimate goal of our framework is to localize effective GNN CL learning (Zhu et al., 2021b) for brain networks. Given a dataset $\mathcal{D}$ and an anchor node i from graph $\mathcal{G}_p \in \mathcal{D}$ with the learned representation $z_{i,p}$, we propose to categorize the possible sample selections into three fundamental types (a visualization is shown in Figure 2):

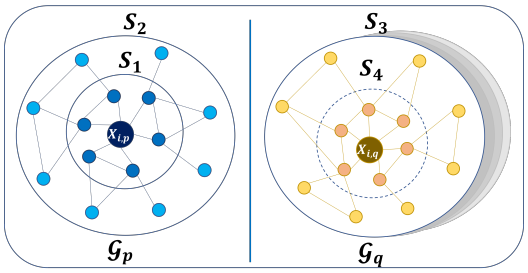


Figure 2: Visual demonstration of the sample types where $X_{i,p}$ is the anchor and $\mathbf{S_1}$/$\mathbf{S_4}$ are sampled as 1-hop neighbors.

- $\underline{\mathbf{S_1}}$: $\{z_{j,p} : j \in \mathcal{N}_k(i,p)\}$ refers to the node representation set within the the k-hop neighborhood of the anchor in graph $\mathcal{G}_p$.
- $\underline{\mathbf{S_2}}$: $\{z_{j,p} : j \notin \mathcal{N}_k(i,p)\}$ refers to the remaining node representation set in graph $\mathcal{G}_p$ that are not in the the k-hop neighborhood of the anchor.
- $\underline{\mathbf{S_3}}$: $\{z_{j,q} : \mathcal{G}_q \in \mathcal{D}, j \in \mathcal{G}_q, q \neq p\}$ refers to the node representation set of nodes in all the other graphs of dataset $\mathcal{D}$.

Notice that our framework leverages the k-hop substructure around the anchor node to further differentiate $\mathbf{S_1}$ and $\mathbf{S_2}$ for contrastive optimization. This design is driven by two considerations: **(1) Regarding GNN learning.** Given that node representations are learned from the information aggregation of its k-hop neighborhood, maximizing the MI of an anchor to its k-hop neighbors naturally enhances lossless message passing of GNN convolutions. **(2) Regarding the uniqueness of brain networks.** Brain networks can be anatomically segmented into smaller neural system modules (Cui et al., 2022b), thus capturing subgraph-level knowledge can provide valuable signals for brain-related analysis.

Building on these three fundamental types of samples, we take advantage of the property of brain networks that ROI identities and orders are fixed across samples to introduce an additional sample type. This encourages the GNN to extract shared substructure knowledge by evaluating the MI of an anchor against its presence in other graphs. Given an anchor representation $z_{i,p}$ of node i from graph $\mathcal{G}_p \in \mathcal{D}$, the novel inter-graph sample type is defined as:

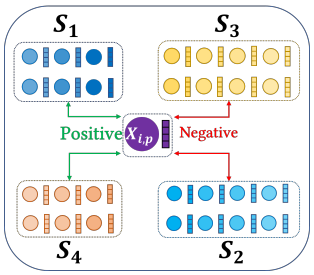


Figure 3: The sampling configuration of the proposed PTGB framework. $\mathbf{S_1}$ and $\mathbf{S_4}$ are positive samples, $\mathbf{S_2}$ and the set $\mathbf{S_3} - \mathbf{S_4}$ are negative samples.

Table 1: The sampling configuration of some existing graph contrastive learning methods. "+" denotes positive sampling, "-" for negative, and "/" for no consideration.

	$\mathbf{S_1}$	$\mathbf{S_2}$	$\mathbf{S_3}$	$\mathbf{S_4}$
DGI	+	+	/	/
InfoG	+	+	–	/
GCC	+	–	–	/
EGI	+	–	–	/
Ours	+	–	–	+

- $\underline{\mathbf{S_4}}$:$\{z_{j,q} : j \in \mathcal{N}_k(i,q) \cap \mathcal{N}_k(i,p), \mathcal{G}_q \in \mathcal{D}, q \neq p\}$, refers to the node representation set within the k-hop neighborhood of node i in all other graphs in $\mathcal{D}$. Conceptually, $\mathbf{S_4}$ is a special subset of $\mathbf{S_3}$.

It is important to note that for an anchor node i, its k-hop neighborhood structures might not be identical among different graphs. As a result, we only consider shared neighborhoods when evaluating the mutual information across multiple graphs. To encourage the learning of unique neighborhood knowledge within a single brain network instance and shared substructure knowledge across the entire dataset, we configure $\mathbf{S_1}$ and $\mathbf{S_4}$ as positive samples while $\mathbf{S_2}$ and the set $\mathbf{S_3} - \mathbf{S_4}$ as negative samples, as illustrated in Figure 3. Strictly speaking, $\mathbf{S_1}$ does not include the anchor itself, but the anchor is always a positive sample to itself by default. Furthermore, our sampling categorization can also help understand the objective formulations in various state-of-the-art graph CL frameworks (Velickovic et al., 2019; Qiu et al., 2020; Xia et al., 2022; Sun et al., 2019; Zhu et al., 2021a). We summarize our findings in Table 1. Specifically, "+" denotes positive sampling; "-" denotes negative sampling; and "/" means that the sample type is not considered. It can be observed that DGI and InfoGraph (InfoG) use graph representation pooled from node representations as a special sample, which is essentially equivalent to jointly considering $\mathbf{S_1}$ and $\mathbf{S_2}$ without explicit differentiation. On the other hand, GCC and EGI, which are more closely related to our framework, leverage neighborhood mutual information maximization on a single graph, but fail to extend this to a multi-graph setting like ours.

3.3. Data-driven Brain Atlas Mapping

Motivation. When fine-tuning a pre-trained model on a new data domain, the misalignment between source and target signals can negatively impact its adaptation. This issue is particularly relevant in brain networks, where it is hard, if not impossible, to require every brain network data provider to stick to the same brain atlas template, and each template can use a unique system of ROIs. For instance, the HIV dataset we obtained is parcellated from the AAL90 template (Tzourio-Mazoyer et al., 2002), leading to 90 defined ROIs; while the PPMI dataset uses the Desikan-Killiany84 template (Desikan et al., 2006), resulting in 84 defined ROIs. As a result, brain networks in the two datasets will have different ROI semantics and graph structures. Although GNNs can handle graphs without fixed numbers and orders of nodes, constructing the most informative ROI (*i.e.,* node) features as the connection profiles (*i.e.,* adjacency) (Cui et al., 2022a,e) can result in different feature dimensions and physical meanings. While manual conversion can be performed to translate between templates, it is a costly process that requires domain expertise to perform even coarse cross-atlas mappings.

To address this issue, we aim to provide a data-driven atlas mapping solution that is easily accessible and eliminates the strong dependency on network construction. The data-driven atlas mapping solution, which transforms the original node features into lower-dimensional representations that preserve the original connectivity information and align features across datasets, is learned independently on each dataset prior to GNN pre-training.

3.3.1. Autoencoder with Brain Network Oriented Regularizers

PTGB adopts a one-layer linear autoencoder (AE) as the base structure. The AE consists of a linear projection encoder $\mathbf{W}$ and a transposed decoder $\mathbf{W}^\top$,

with the goal of learning a low-dimensional projection that can easily reconstruct the original presentation. The loss function is defined as minimizing the reconstruction error $\mathcal{L}_{\text{rec}} = (1/M)\|\mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{W}^\top\|_2^2$, where $\mathbf{X} \in \mathbb{R}^{M\times M}$ is the input and $\mathbf{W} \in \mathbb{R}^{M\times D}$ is the learnable projection (Hinton and Zemel, 1993). To further enhance the feature compression and to guide the overall AE optimization, we propose to incorporate several regularizers that take into account the unique characteristics of brain networks.:

Locality-Preserving Regularizer (LR). We aim to ensure that the compressed features preserve the spatial relationships of the original brain surface. To achieve this, we incorporate a locality preserving regularizer (He et al., 2005) to the AE objective. The regularizer is formulated as $\mathcal{L}_{\text{loc}} = (1/M)\|\mathbf{Y}-\mathbf{T}\mathbf{Y}\|^2$, where $\mathbf{Y} \in \mathbb{R}^{M\times D}$ represents the projected features from the AE and $\mathbf{T} \in \mathbb{R}^{M\times M}$ is a transition matrix constructed from the k-NN graph of the 3D coordinates of ROIs.

Modularity-Aware Regularizer (CR). Brain networks can be segmented into various neural system modules that characterize functional subsets of ROIs. In graph terminology, they are community structures. The projected feature should also capture information about neural system membership. However, obtaining ground-truth segmentations is a difficult task that requires expert knowledge. To overcome this challenge, we resort to community detection methods on graphs, specifically based on modularity maximization. The regularizer (Salha-Galvan et al., 2022) is defined as minimizing

$$\mathcal{L}_{\text{com}} = -\frac{1}{2D}\sum_{i,j=1}^{M}\left[\mathbf{A}_{ij} - \frac{k_i k_j}{2D}\right]\exp(-\|y_i - y_j\|_2^2), \tag{5}$$

where $\mathbf{A} \in \mathbb{R}^{M\times M}$ is the graph adjacency matrix, k_i denotes degree of node i, and y_i is the AE projected features. Essentially, this optimization minimizes the L_2 distance between representations of nodes within the same communities, as measured by the modularity score, and maximizes the distance between representations of nodes in different communities.

Sparsity-Oriented Regularizer (SC). Sparse networks have proven to be effective in learning robust representations from noisy data (Jeong et al., 2017; Shi et al., 2019; Makhzani and Frey, 2014). In brain connectome analysis, sparsity has also been shown to improve the interpretation of task-specific ROI connections in generation and classification tasks (Kan et al., 2022a). To this end, we implement the popular KL-divergence smoothing to enforce sparsity in the parameters of the linear projection encoder, $\mathbf{W}$). This is formulated as:

$$\mathcal{L}_{\text{KL}} = \sum_{i=1}^{M}\sum_{j=1}^{D}\left[\rho\log\left(\frac{\rho}{\hat{\rho}_{ij}}\right) + (1-\rho)\log\left(\frac{1-\rho}{1-\hat{\rho}_{ij}}\right)\right], \tag{6}$$

where ρ is a small positive float set as the target sparsity value, and $\hat{\rho}_{ij}$ represents the element-wise activation of the encoder projection matrix $\mathbf{W} \in \mathbb{R}^{M\times D}$.

3.3.2. Variance-based Dimension Sorting

In addition to transforming dataset-specific features, cross-dataset alignment of feature signals is also crucial for improving model adaptation. The one-layer AE transforms the original feature vectors into weighted combinations of multiple dimensions, creating new feature dimensions which we name as *virtual ROIs*. In the context of brain networks, this process helps to group ROIs and their signals. This idea is inspired by the well-studied functional brain modules (Philipson, 2002; Anderson et al., 2004; Hilger et al., 2020; Brodmann, 1909; Zhou et al., 2020), which provide a higher-level and generic organization of the brain surface, as opposed to fine-grained ROI systems. Since the variations in ROI parcellations are due to differences in clinical conventions, it is reasonable to assume that there exists a shared virtual ROI system underlying different parcellation systems, similar to the discretization of functional brain modules. The community learning and neighborhood preserving regularizers, introduced in Section 3.3, allow us to capture these shared virtual ROIs in a data-driven manner. Our ultimate goal is to align the discovered virtual ROIs across datasets, so that each virtual ROI characterizes the same functional module in the human brain, regardless of its origin. This cross-dataset alignment of virtual ROIs ensures that the model can effectively adapt to new datasets and provide meaningful insights into the different downstream analyses.

The objective of the one-layer linear AE is similar to PCA, as discussed in more detail in Appendix A.1, with the added benefit of incorporating additional regularizers. PCA orders dimensions based on decreasing levels of sample variance (Hotelling, 1933). PTGB leverage this approach by utilizing the learned parameters of the AE projection to estimate the variance of each virtual ROI (*i.e.*, projected feature di-

mension). The sample variance of each virtual ROI indicates its representativeness of the original data variations. Given the shared patterns across different parcellation systems, we expect that similar virtual ROIs in datasets with different atlas templates will have similar variance scores, especially in terms of their order. By sorting the same number of virtual ROIs based on their sample variance in each dataset, we aim to align virtual ROI cross datasets, so that each virtual ROI represents the same functional unit in the human brain. The procedure is explained in detail in Algorithm 2 in Appendix A.2.

4. Experiments

We evaluate the effectiveness of PTGB through extensive experiments on real brain network datasets, with a focus on the following research questions:

- **RQ1**: How does PTGB compare with other unsupervised GNN pre-training frameworks adapted to the scenario of brain networks?
- **RQ2**: What is the contribution of each major component in PTGB to the overall performance?
- **RQ3**: How does the choice of sampling method affect model convergence and performance?
- **RQ4**: How effective is the variance-based sorting in aligning virtual ROIs among different parcellation systems?

Datasets, Configurations, and Metrics. Our experiments are conducted on three real-world brain network datasets: PPMI, BP, and HIV. The PPMI dataset is parcellated using the Desikan-Killiany84 atlas template and includes brain networks from 718 subjects, 569 of whom are Parkinson's Disease (PD) patients and 149 are Healthy Control (HC). The networks are constructed using three tractography algorithms: Probabilistic Index of Connectivity (PICo), Hough voting (Hough), and FSL. The BP dataset is parcellated using the Brodmann82 template and includes resting-state fMRI and DTI modalities from 97 subjects, 52 of whom have Bipolar I disorder and 45 are HCs. The HIV dataset is parcellated using the AAL90 template and includes fMRI and DTI modalities from 70 subjects, with 35 early HIV patients and 35 HCs. We pre-train the model on the PPMI dataset and evaluate the downstream performance on BP and HIV. Further details about the datasets can be found in Appendix B.

PTGB employs GCN as the backbone for the GNN (Kipf and Welling, 2017) encoder. We also benchmark PTGB with GAT (Veličković et al., 2018) and GIN (Xu et al., 2019), and the results are provided in Appendix D.1. The hyperparameter settings are described in detail in Appendix C. The hyperparameter tuning follows the standard designs in related studies such as in (Yang et al., 2021; Wein et al., 2021; Hu et al., 2021). The downstream evaluation is binary graph classification for disease prediction. To assess the performance, we use the two widely used metrics in the medical field (Li et al., 2021a; Cui et al., 2022a): accuracy score (ACC) and the area under the receiver operating characteristic curve (AUC).

4.1. Overall Performance Comparison (RQ1)

We present a comprehensive comparison of the target performance between the proposed PTGB and popular unsupervised learning strategies in Table 2. To fairly compare the methods, we apply atlas mapping pre-processing and the multi-dataset learning backbone discussed in section 3.1 to all methods. The purpose of this comparison is to effectively highlight the impact of the proposed two-level contrastive pre-training and we will further analyze the effect of atlas mapping in subsequent subsections. In addition, for a clearer presentation, we group the selected baselines according to their optimization strategies:

- No pre-training (NPT): the backbone with randomly initialized parameters for target evaluation.
- Non-CL-based (NCL): methods with cost functions regularized by co-occurrence agreement or link reconstruction, including Node2Vec (Grover and Leskovec, 2016), DeepWalk (Perozzi et al., 2014), and VGAE (Kipf and Welling, 2016).
- Single-scale CL (SCL): methods utilizing either node- or graph-level representations in the CL optimization, including GBT (Bielak et al., 2022), ProGCL (Xia et al., 2022), and GraphCL (You et al., 2020).
- Multi-scale CL (MCL): methods whose CL optimization utilizes both nodes- and graph-level representations, including DGI (Velickovic et al., 2019) and InfoG (Sun et al., 2019).
- Ego-graph sampling (EGS): methods whose contrastive samplings consider k-hop ego-networks as discriminative instances, which are the most similar to the proposed PTGB, including GCC (Qiu et al., 2020) and EGI (Zhu et al., 2021a).
- Our proposed two-level contrastive optimization (Ours): methods include single task pre-training (STP) in which we select the PICo modality of

Table 2: Disease prediction performance comparison. All results are averaged from 5-fold cross-validation along with standard deviations. The best result is highlighted in bold and runner-up is underlined. * denotes a significant improvement according to paired t-test with $\alpha = 0.05$ compared with baselines.

Type	Method	BP-fMRI		BP-DTI		HIV-fMRI		HIV-DTI	
		ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
NPT	GCN	$50.07_{\pm 13.70}$	$50.11_{\pm 15.48}$	$49.51_{\pm 14.68}$	$51.83_{\pm 13.98}$	$56.27_{\pm 15.84}$	$57.16_{\pm 15.14}$	$51.30_{\pm 16.42}$	$53.82_{\pm 14.94}$
NCL	Node2Vec	$48.51_{\pm 10.39}$	$49.68_{\pm 7.23}$	$50.83_{\pm 8.14}$	$46.70_{\pm 10.33}$	$52.61_{\pm 10.38}$	$50.75_{\pm 10.94}$	$49.65_{\pm 10.30}$	$51.22_{\pm 10.79}$
	DeepWalk	$50.28_{\pm 9.33}$	$51.59_{\pm 9.06}$	$52.17_{\pm 9.74}$	$48.36_{\pm 9.37}$	$54.81_{\pm 11.26}$	$55.55_{\pm 11.93}$	$52.67_{\pm 11.42}$	$50.88_{\pm 10.53}$
	VGAE	$56.71_{\pm 9.68}$	$55.24_{\pm 11.48}$	$54.63_{\pm 12.09}$	$54.21_{\pm 11.94}$	$62.76_{\pm 9.47}$	$61.25_{\pm 11.61}$	$56.90_{\pm 9.72}$	$55.35_{\pm 9.04}$
SCL	GBT	$57.21_{\pm 10.68}$	$57.32_{\pm 10.48}$	$56.29_{\pm 9.35}$	$55.27_{\pm 10.54}$	$65.73_{\pm 10.93}$	$66.08_{\pm 10.43}$	$59.80_{\pm 9.76}$	$57.37_{\pm 9.49}$
	GraphCL	$59.79_{\pm 9.36}$	$59.10_{\pm 10.78}$	$57.57_{\pm 10.63}$	$57.35_{\pm 9.67}$	$67.08_{\pm 9.70}$	$69.17_{\pm 10.68}$	$60.43_{\pm 8.39}$	$60.03_{\pm 10.48}$
	ProGCL	$62.36_{\pm 8.90}$	$62.61_{\pm 9.34}$	$61.26_{\pm 8.37}$	$\underline{62.67_{\pm 8.46}}$	$71.52_{\pm 9.19}$	$72.16_{\pm 9.85}$	$\underline{62.48_{\pm 10.38}}$	$61.94_{\pm 10.57}$
MCL	DGI	$62.44_{\pm 10.12}$	$60.75_{\pm 10.97}$	$58.15_{\pm 9.63}$	$58.95_{\pm 9.60}$	$70.22_{\pm 11.43}$	$70.12_{\pm 12.46}$	$60.83_{\pm 10.84}$	$62.06_{\pm 10.16}$
	InfoG	$62.87_{\pm 9.52}$	$62.37_{\pm 9.67}$	$60.88_{\pm 9.97}$	$60.44_{\pm 9.61}$	$72.46_{\pm 8.71}$	$72.94_{\pm 8.68}$	$61.75_{\pm 9.76}$	$61.37_{\pm 9.85}$
EGS	GCC	$\underline{63.45_{\pm 9.82}}$	$62.39_{\pm 9.08}$	$60.44_{\pm 9.54}$	$60.29_{\pm 10.33}$	$70.97_{\pm 10.31}$	$72.48_{\pm 11.36}$	$61.27_{\pm 9.66}$	$61.38_{\pm 10.72}$
	EGI	$63.38_{\pm 8.93}$	$\underline{63.58_{\pm 8.02}}$	$\underline{61.82_{\pm 8.53}}$	$61.57_{\pm 8.27}$	$\underline{73.46_{\pm 8.49}}$	$\underline{73.28_{\pm 8.68}}$	$60.89_{\pm 9.87}$	$\underline{62.41_{\pm 8.50}}$
Ours	STP	$53.92_{\pm 12.82}$*	$54.61_{\pm 11.76}$*	$55.51_{\pm 15.74}$*	$56.73_{\pm 16.23}$*	$61.18_{\pm 14.57}$*	$62.88_{\pm 15.58}$*	$55.29_{\pm 12.38}$*	$57.31_{\pm 14.72}$*
	MTP	$60.37_{\pm 12.42}$*	$61.64_{\pm 11.83}$*	$59.41_{\pm 11.62}$*	$59.92_{\pm 13.37}$*	$67.65_{\pm 12.26}$*	$68.38_{\pm 12.94}$*	$60.54_{\pm 13.83}$*	$59.46_{\pm 12.33}$*
	PTGB	$\mathbf{68.84_{\pm 8.26}}$*	$\mathbf{68.45_{\pm 8.96}}$*	$\mathbf{66.57_{\pm 7.67}}$*	$\mathbf{68.31_{\pm 9.39}}$*	$\mathbf{77.80_{\pm 9.76}}$*	$\mathbf{77.22_{\pm 8.74}}$*	$\mathbf{67.51_{\pm 8.67}}$*	$\mathbf{67.74_{\pm 8.59}}$*

the PPMI study to be the only source task; multi-task pre-trainig (MTP) which does not utilize the MAML technique; and the full implementation of the PTGB framework.

The experiments reveal the following insights:

- The proposed PTGB consistently outperforms all the baselines, achieving a relative improvement of 7.34%-13.30% over the best-performing baselines and 31.80%-38.26% over the NPT setting. The results of PTGB have been statistically compared against baselines using paired t-tests. With a significance level set to 0.05, the largest two-tailed p value is reported at 0.042, indicating that PTGB demonstrates a statistically significant performance increase over other selected methods.
- Compared with the transductive methods of Node2Vec and DeepWalk, the GNN pre-trained by VGAE learns structure-preserving representations and achieves the best results in the NCL-type methods. This indicates the potential benefit of the locality-preserving regularizer design in PTGB.
- Maximizing mutual information between augmented instances may hinder GNNs from learning a shared understanding of the entire dataset. For baselines belonging to the categories of SCL, MCL, and EGS, pre-training with non-augmented CL (InfoG, EGI) generally results in a 4.36% relative improvement across both metrics and a 7.63% relative decrease in performance variance compared to their augmentation-based counterparts (GBT, GraphCL, ProGCL, DGI, GCC). This explains why PTGB does not employ data augmentation.
- Multi-scale MI promotes the capture of effective local (*i.e.,* node-level) representations that can summarize the global (*i.e.,* graph-level) information of the entire network. The MCL-type methods typically outperform the SCL-type ones by a relative gain of 2.68% in ACC and 3.27% in AUC.
- The group of baselines considering k-hop neighborhoods (EGS) presents the strongest performance, indicating the importance of local neighborhoods in brain network analysis. The proposed PTGB, which captures this aspect through both node- and graph-level CL, is the only one that comprehensively captures the local neighborhoods of nodes.
- Learning from multiple tasks (MTP) brings significant improvement over STP, reporting a relative increase of 8.47% in accuracy and 6.90% in AUC. Furthermore, the full PTGB framework with MAML-styled training achieves a relative improvement of 11.29% in accuracy, 14.75% in AUC, and a reduced variance over MTP, demonstrating its advantages in enhancing model generalizability.

4.2. Ablation Studies (RQ2)

We examine two key components of PTGB- (1) the two-level contrastive sampling and (2) the atlas mapping regularizers. The best contrastive sampling configuration is fixed when examining the atlas regularizers, and all regularizers are equipped when examin-

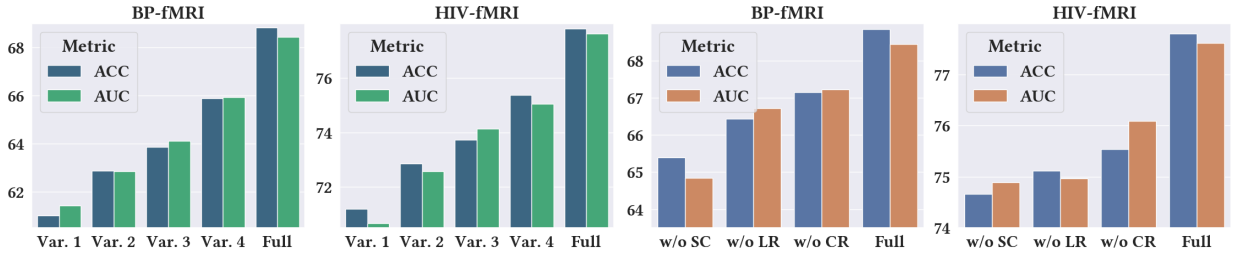
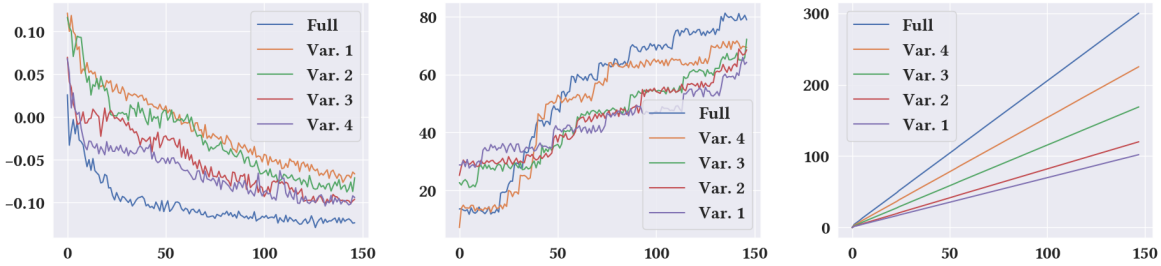


Figure 4: Ablation comparisons on contrastive sampling choices (left two) and atlas mapping regularizers (right two). The y-axis refers to the numeric values of evaluated metrics (in %). The setup of Var. 1 - 4 is described in Table 3. "SC", "LR", and "CR" are abbreviations for "sparsity constraints", "locality regularizer", and "community (modularity-aware) regularizer" respectively.



(a) Pre-train loss on PPMI (b) Test ACC (%) on HIV-fMRI (c) Pre-train time (secs) on PPMI

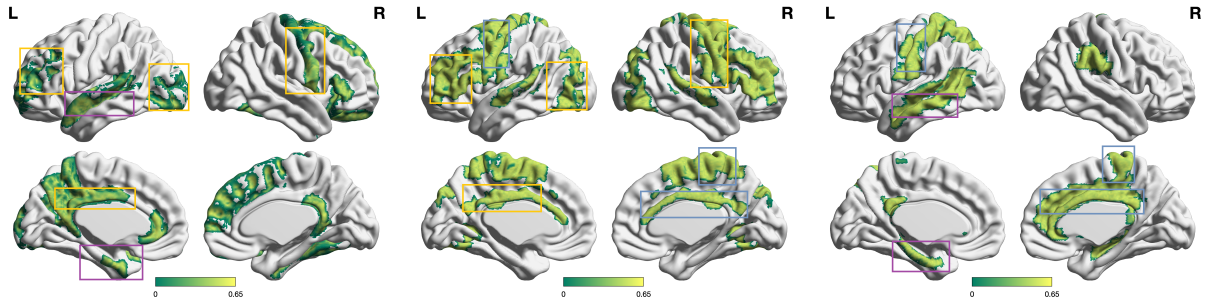
Figure 5: In-depth comparison among the four variants and the full model. The x-axis is epochs. Fig. (a) evaluates the trajectory of pre-training loss, Fig. (b) evaluates their respective testing accuracy on the fMRI view of the HIV dataset, and Fig. (c) reports the pre-training runtime in seconds.

Table 3: The four variants of sampling strategies.

	$\mathbf{S_1}$	$\mathbf{S_2}$	$\mathbf{S_3}$	$\mathbf{S_4}$
Var. 1	–	–	/	/
Var. 2	+	–	/	/
Var. 3	+	–	–	/
Var. 4	+	+	–	/

ing the contrastive samplings. The results, shown in Figure 4 (with additional DTI version in Appendix D.2), are analyzed based on the four possible variants of contrastive sampling listed in Table 3. Our analyses yield the following observations: **(1)** leveraging k-hop neighborhood (*i.e.,* positive $\mathbf{S_2}$) MI maximization brings visible performance gain, confirming its benefit in brain structure learning; **(2)** The extension to multi-graph CL (*i.e.,* consideration of $\mathbf{S_3}$) facilitates the extraction of unique ROI knowledge, leading to improved results in Var. 3/4; **(3)** Var. 4 outperforms Var. 3 as it effectively summarizes of global (*i.e.,* graph-level) information in local node representations; **(4)** The full implementation of PTGB brings a relative gain of 4.27% in both metrics on top of Var. 4, highlighting the significance of considering shared substructure knowledge across multiple graphs (*i.e.,* through the inclusion of $\mathbf{S_4}$).

The right-side sub-figures examine the impact of the atlas mapping regularizers by comparing the results of the full framework to those without the sparsity regularizer (w/o SR), the locality regularizer (w/o LR), and the community regularizer (w/o CR). Two key observations are made: **(1)** The removal of SR leads to the greatest performance drop, emphasizing its crucial role in learning robust projections that can effectively handle noise and prevent over-fitting; **(2)** The inferior results when LR and CR are absent emphasize the importance of spatial sensitivity and blockwise feature information in brain network analysis. This supports our intuition to consider the relative positioning of ROIs in the 3D coordinate as



(a) PPMI mapping (b) BP mapping (c) HIV mapping

Figure 6: The virtual ROI mapping across the three investigated datasets. We highlight pairs of overlapping regions with colored boxes. In particular, we use gold boxes for the PPMI and BP mapping; blue boxes for the BP and HIV mapping; and purple boxes for the PPMI and HIV mapping.

well as knowledge on community belongings based on modularity measures.

4.3. Analysis of Two-level Contrastive Sampling (RQ3)

Figure 5 offers insight into the pre-training convergence, target adaptation progression, and pre-training runtime consumption of the four sampling variants and the full framework. Key observations include: **(1)** As seen in Figure 5(a), all variants demonstrate efficient pre-training convergence due to the multi-dataset joint optimization inspired by MAML. The full model demonstrates the most optimal convergence, highlighting the advantage of learning shared neighborhood information in brain network data through two-level node contrastive sampling. **(2)** Figure 5(b) shows the superiority of our design in terms of downstream adaptation performance compared to other variants. **(3)** Figure 5(c) reveals that the more sophisticated the sampling considerations result in greater computational complexity for mutual information evaluation, leading to longer runtime for each pre-training epoch. However, the total time consumptions are all on the same scale.

4.4. Analysis of ROI Alignment (RQ4)

To further validate the variance-based virtual ROI sorting, we select the top 2 virtual ROIs with the highest sample variances for each atlas template (*i.e.,* dataset) and backtrack to locate their corresponding projected ROIs. The results are illustrated in Figure 6, which shows a 3D brain surface visualization highlighting the original ROIs. From this, we draw two main conclusions: **(1)** There exists multiple regional overlaps between pairs of two atlas templates, reflecting some working effectiveness of our proposed solution as well as confirming the feasibility of converting between atlas templates. **(2)** It is relatively harder to find regions that overlap across all three atlas templates which shows a limitation of the proposed unsupervised ROI alignment scheme, suggesting a need to modify against the current variance-based heuristic which may inspire further study and research opportunity.

5. Conclusion

Brain network analysis for task-specific disease prediction has been a challenging task for conventional GNN frameworks due to the limited availability of labeled training data and the absence of a unifying brain atlas definition, which hinders efficient knowledge transfer across different datasets. To address these challenges, we propose PTGB, a novel unsupervised multi-dataset GNN pre-training that leverages a two-level node contrastive sampling to overcome data scarcity. Additionally, PTGB incorporates atlas mapping through brain-network-oriented regularizers and variance-based sorting to address the issue of incompatible ROI parcellation systems in cross-dataset model adaptation in a data-driven way. Extensive experiments on real-world brain connectome datasets demonstrate the superiority and robustness of PTGB in disease prediction and its clear advantage over various state-of-the-art baselines. As more brain network datasets become available, it will be intriguing to further validate its generalizability.

References

Darko Aleksovski, Dragana Miljkovic, Daniele Bravi, and Angelo Antonini. Disease progression in parkinson subtypes: the ppmi dataset. *Neurol. Sci.*, 39:1971–1976, 2018.

John R Anderson, Daniel Bothell, Michael D Byrne, Scott Douglass, Christian Lebiere, and Yulin Qin. An integrated theory of the mind. *Psychological review*, 2004.

Jimmy Ba, Geoffrey E Hinton, Volodymyr Mnih, Joel Z Leibo, and Catalin Ionescu. Using fast weights to attend to the recent past. *NeurIPS*, 2016.

Piotr Bielak, Tomasz Kajdanowicz, and Nitesh V Chawla. Graph barlow twins: A self-supervised representation learning framework for graphs. *Knowl Based Syst*, 2022.

Korbinian Brodmann. *Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues*. 1909.

Bokai Cao, Liang Zhan, Xiangnan Kong, Philip S Yu, Nathalie Vizueta, Lori L Altshuler, and Alex D Leow. Identification of discriminative subgraph patterns in fmri brain networks in bipolar affective disorder. In *International Conference on Brain Informatics and Health*, 2015.

Yue Cao, Zhenda Xie, Bin Liu, Yutong Lin, Zheng Zhang, and Han Hu. Parametric instance classification for unsupervised visual feature learning. *NeurIPS*, 2020.

BJ 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. *Dev. Cogn. Neurosci.*, 32: 43–54, 2018.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020a.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020b.

Hejie Cui, Wei Dai, Yanqiao Zhu, Xuan Kan, Antonio Aodong Chen Gu, Joshua Lukemire, Liang Zhan, Lifang He, Ying Guo, and Carl Yang. Braingb: A benchmark for brain network analysis with graph neural networks. *IEEE TMI*, 2022a.

Hejie Cui, Wei Dai, Yanqiao Zhu, Xiaoxiao Li, Lifang He, and Carl Yang. Interpretable graph neural networks for connectome-based brain disorder analysis. 2022b.

Hejie Cui, Wei Dai, Yanqiao Zhu, Xiaoxiao Li, Lifang He, and Carl Yang. Interpretable graph neural networks for connectome-based brain disorder analysis. In *MICCAI*, 2022c.

Hejie Cui, Jiaying Lu, Yao Ge, and Carl Yang. How can graph neural networks help document retrieval: A case study on cord19 with concept map generation. In *ECIR*, 2022d.

Hejie Cui, Zijie Lu, Pan Li, and Carl Yang. On positional and structural node features for graph neural networks on non-attributed graphs. In *CIKM*, 2022e.

Wei Dai, Hejie Cui, Xuan Kan, Ying Guo, and Carl Yang. Transformer-based hierarchical clustering for brain network analysis. In *ISBI*, 2023.

Rahul S Desikan, Florent Ségonne, Bruce Fischl, Brian T Quinn, Bradford C Dickerson, Deborah Blacker, Randy L Buckner, Anders M Dale, R Paul Maguire, Bradley T Hyman, et al. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, 2006.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

Adriana Di Martino, Chao-Gan Yan, Qingyang Li, Erin Denio, Francisco X Castellanos, Kaat Alaerts, Jeffrey S Anderson, Michal Assaf, Susan Y Bookheimer, Mirella Dapretto, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry*, 2014.

Federico Errica, Marco Podda, Davide Bacciu, and Alessio Micheli. A fair comparison of graph neural networks for graph classification. In *ICLR*, 2020.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *NeurIPS*, 2020.

Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *SIGKDD*, 2016.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *NeurIPS*, 2017.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

Xiaofei He, Deng Cai, Shuicheng Yan, and Hong-Jiang Zhang. Neighborhood preserving embedding. In *ICCV*, 2005.

Kirsten Hilger, Makoto Fukushima, Olaf Sporns, and Christian J Fiebach. Temporal stability of functional brain modules associated with human intelligence. *Human brain mapping*, 2020.

Chris Hinrichs, Vikas Singh, Lopamudra Mukherjee, Guofan Xu, Moo K Chung, Sterling C Johnson, Alzheimer's Disease Neuroimaging Initiative, et al. Spatially augmented lpboosting for ad classification with evaluations on the adni dataset. *NeuroImage*, 48:138–149, 2009.

Geoffrey E Hinton and Richard Zemel. Autoencoders, minimum description length and helmholtz free energy. *NeurIPS*, 1993.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *ICLR*, 2019.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *J Educ Psychol*, 1933.

Yupeng Hou, Binbin Hu, Wayne Xin Zhao, Zhiqiang Zhang, Jun Zhou, and Ji-Rong Wen. Neural graph matching for pre-training graph neural networks. In *SDM*, 2022.

Jinlong Hu, Lijie Cao, Tenghui Li, Shoubin Dong, and Ping Li. Gat-li: a graph attention network based learning and interpreting method for functional brain network classification. *BMC bioinformatics*, 2021.

W Hu, B Liu, J Gomes, M Zitnik, P Liang, V Pande, and J Leskovec. Strategies for pre-training graph neural networks. In *ICLR*, 2020a.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. 2020b.

Ziniu Hu, Yuxiao Dong, Kuansan Wang, Kai-Wei Chang, and Yizhou Sun. Gpt-gnn: Generative pre-training of graph neural networks. In *SIGKDD*, 2020c.

Seongah Jeong, Xiang Li, Jiarui Yang, Quanzheng Li, and Vahid Tarokh. Dictionary learning and sparse coding-based denoising for high-resolution task functional connectivity mri analysis. In *International Workshop on Machine Learning in Medical Imaging*, 2017.

Xuan Kan, Hejie Cui, Joshua Lukemire, Ying Guo, and Carl Yang. Fbnetgen: Task-aware gnn-based fmri analysis via functional brain network generation. *MIDL*, 2022a.

Xuan Kan, Wei Dai, Hejie Cui, Zilong Zhang, Ying Guo, and Carl Yang. Brain network transformer. In *NeurIPS*, 2022b.

Xuan Kan, Yunchuan Kong, Tianwei Yu, and Ying Guo. Bracenet: Graph-embedded neural network for brain network analysis. In *IEEE Big Data*, 2022c.

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

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Thomas N Kipf and Max Welling. Variational graph auto-encoders. *NIPS Workshop on Bayesian Deep Learning*, 2016.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

Xiaoxiao Li, Yuan Zhou, Nicha Dvornek, Muhan Zhang, Siyuan Gao, Juntang Zhuang, Dustin Scheinost, Lawrence H Staib, Pamela Ventola, and James S Duncan. Braingnn: Interpretable brain graph neural network for fmri analysis. *Med Image Anal*, 2021a.

Xiaoxiao Li, Yuan Zhou, Nicha C. Dvornek, Muhan Zhang, Siyuan Gao, Juntang Zhuang, Dustin Scheinost, Lawrence H. Staib, Pamela Ventola, and James S. Duncan. Braingnn: Interpretable brain graph neural network for fmri analysis. *Med. Image Anal.*, 2021b.

Chen Liang, Yue Yu, Haoming Jiang, Siawpeng Er, Ruijia Wang, Tuo Zhao, and Chao Zhang. Bond: Bert-assisted open-domain named entity recognition with distant supervision. In *KDD*, 2020.

Martin A Lindquist. The statistical analysis of fmri data. *Stat Sci*, 23:439–464, 2008.

Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. Pre-training molecular graph representation with 3d geometry. *ICLR*, 2022.

Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *ICLR*, 2017.

Yuanfu Lu, Xunqiang Jiang, Yuan Fang, and Chuan Shi. Learning to pre-train graph neural networks. In *AAAI*, 2021.

Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. Parameterized explainer for graph neural network. In *NeurIPS*, 2020.

Gongxu Luo, Chenyang Li, Hejie Cui, Lichao Sun, Lifang He, and Carl Yang. Multi-view brain network analysis with cross-view missing network generation. In *BIBM*, 2022.

Guixiang Ma, Chun-Ta Lu, Lifang He, S Yu Philip, and Ann B Ragin. Multi-view graph embedding with hub detection for brain network analysis. In *ICDM*, 2017.

Alireza Makhzani and Brendan Frey. K-sparse autoencoders. *ICLR*, 2014.

Gustav Martensson, Joana B Pereira, Patrizia Mecocci, Bruno Vellas, Magda Tsolaki, Iwona Kłoszewska, Hilkka Soininen, Simon Lovestone, Andrew Simmons, Giovanni Volpe, et al. Stability of graph theoretical measures in structural brain networks in alzheimer's disease. *Scientific reports*, 8:1–15, 2018.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *SIGKDD*, 2014.

Lars Philipson. Functional modules of the brain. *Journal of theoretical biology*, 2002.

Jiezhong Qiu, Qibin Chen, Yuxiao Dong, Jing Zhang, Hongxia Yang, Ming Ding, Kuansan Wang, and Jie Tang. Gcc: Graph contrastive coding for graph neural network pre-training. In *SIGKDD*, 2020.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of maml. 2019.

Ryan A Rossi and Nesreen K Ahmed. An interactive data repository with visual analytics. *SIGKDD*, 2016.

Guillaume Salha-Galvan, Johannes F. Lutzeyer, George Dasoulas, Romain Hennequin, and Michalis Vazirgiannis. Modularity-aware graph autoencoders for joint community detection and link prediction. *Neural Netw*, 2022.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *ESWC*, 2018.

Rui Shi, Jian Ji, Chunhui Zhang, and Qiguang Miao. Boosting sparsity-induced autoencoder: A novel sparse feature ensemble learning for image classification. *Int. J. Adv. Robot. Syst.*, 2019.

Stephen M Smith. The future of fmri connectivity. *NeuroImage*, 62:1257–1266, 2012.

Fan-Yun Sun, Jordan Hoffman, Vikas Verma, and Jian Tang. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. In *ICLR*, 2019.

Mingchen Sun, Kaixiong Zhou, Xin He, Ying Wang, and Xin Wang. Gppt: Graph pre-training and prompt tuning to generalize graph neural networks. In *SIGKDD*, 2022.

Hong Hui Tan and King Hann Lim. Review of second-order optimization techniques in artificial neural networks backpropagation. *IOP Conference Series: Materials Science and Engineering*, 2019.

Nathalie Tzourio-Mazoyer, Brigitte Landeau, Dimitri Papathanassiou, Fabrice Crivello, Olivier Etard, Nicolas Delcroix, Bernard Mazoyer, and Marc Joliot. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *NeuroImage*, 2002.

Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha Talukdar. Composition-based multi-relational graph convolutional networks. In *ICLR*, 2020.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *ICLR*, 2018.

Petar Velickovic, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax. *ICLR*, 2019.

M Vu and MT Thai. Pgm-explainer: Probabilistic graphical model explanations for graph neural networks. In *NeurIPS*, 2020.

Simon Wein, Wilhelm M Malloni, Ana Maria Tomé, Sebastian M Frank, G-I Henze, Stefan Wüst, Mark W Greenlee, and Elmar W Lang. A graph neural network framework for causal inference in brain networks. *Scientific reports*, 2021.

Jun Xia, Lirong Wu, Ge Wang, Jintao Chen, and Stan Z Li. Progcl: Rethinking hard negative mining in graph contrastive learning. In *ICML*, 2022.

Wenhan Xiong, Jingfei Du, William Yang Wang, and Veselin Stoyanov. Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model. *ICLR*, 2020.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *ICLR*, 2019.

Ran Xu, Yue Yu, Chao Zhang, Mohammed K Ali, Joyce C Ho, and Carl Yang. Counterfactual and factual reasoning over hypergraphs for interpretable clinical predictions on ehr. In *Machine Learning for Health*, 2022.

Ran Xu, Yue Yu, Hejie Cui, Xuan Kan, Yanqiao Zhu, Joyce Ho, Chao Zhang, and Carl Yang. Neighborhood-regularized self-training for learning with few labels. *AAAI*, 2023a.

Ran Xu, Yue Yu, Joyce C Ho, and Carl Yang. Weakly-supervised scientific document classification via retrieval-augmented multi-stage training. In *the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023b.

Noriaki Yahata, Jun Morimoto, Ryuichiro Hashimoto, Giuseppe Lisi, Kazuhisa Shibata, Yuki Kawakubo, Hitoshi Kuwabara, Miho Kuroda, Takashi Yamada, Fukuda Megumi, et al. A small number of abnormal brain connections predicts adult autism spectrum disorder. *Nat. Commun.*, 7:1–12, 2016.

Chunde Yang, Panyu Wang, Jia Tan, Qingshui Liu, and Xinwei Li. Autism spectrum disorder diagnosis using graph attention network based on spatial-constrained sparse functional brain networks. *Computers in Biology and Medicine*, 2021.

Yi Yang, Yanqiao Zhu, Hejie Cui, Xuan Kan, Lifang He, Ying Guo, and Carl Yang. Data-efficient brain connectome analysis via multi-task meta-learning. In *SIGKDD*, 2022.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *ICLR*, 2022.

Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. In *NeurIPS*, 2019.

Zhitao Ying, Jiaxuan You, Christopher Morris, Xiang Ren, Will Hamilton, and Jure Leskovec. Hierarchical graph representation learning with differentiable pooling. In *NeurIPS*, 2018.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. In *NeurIPS*, 2020.

Yue Yu, Simiao Zuo, Haoming Jiang, Wendi Ren, Tuo Zhao, and Chao Zhang. Fine-tuning pre-trained language model with weak supervision: A contrastive-regularized self-training approach. In *NAACL*, 2021.

Yue Yu, Chenyan Xiong, Si Sun, Chao Zhang, and Arnold Overwijk. Coco-dr: Combating distribution shifts in zero-shot dense retrieval with contrastive and distributionally robust learning. In *EMNLP*, 2022.

Yue Yu, Xuan Kan, Hejie Cui, Ran Xu, Yujia Zheng, Xiangchen Song, Yanqiao Zhu, Kun Zhang, Razieh Nabi, Ying Guo, et al. Learning task-aware effective brain connectivity for fmri analysis with graph neural networks. In *ISBI*, 2023.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *ICML*, 2021.

Shijie Zhang, Hongzhi Yin, Tong Chen, Quoc Viet Nguyen Hung, Zi Huang, and Lizhen Cui. Gcn-based user representation learning for unifying robust recommendation and fraudster detection. In *SIGIR*, 2020.

Tong Zhao, Yozen Liu, Leonardo Neves, Oliver Woodford, Meng Jiang, and Neil Shah. Data augmentation for graph neural networks. In *AAAI*, 2021.

Zhen Zhou, Xiaobo Chen, Yu Zhang, Dan Hu, Lishan Qiao, Renping Yu, Pew-Thian Yap, Gang Pan, Han Zhang, and Dinggang Shen. A toolbox for brain network construction and classification (brainnetclass). *Hum Brain Mapp*, 2020.

Qi Zhu, Carl Yang, Yidan Xu, Haonan Wang, Chao Zhang, and Jiawei Han. Transfer learning of graph neural networks with ego-graph information maximization. In *NeurIPS*, 2021a.

Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Graph Contrastive Learning with Adaptive Augmentation. In *WWW*, 2021b.

Yanqiao Zhu, Hejie Cui, Lifang He, Lichao Sun, and Carl Yang. Joint embedding of structural and functional brain networks with graph neural networks for mental illness diagnosis. In *EMBC*, 2022a.

Yanqiao Zhu, Yichen Xu, Hejie Cui, Carl Yang, Qiang Liu, and Shu Wu. Structure-enhanced heterogeneous graph contrastive learning. In *SDM*, 2022b.

Appendix A. Autoencoder Structure Analysis

A.1. Bridging Reconstruction Minimization and Variance Maximization

In this subsection, we briefly discuss how the reconstruction minimizing objective in one-layer AE can be cast to a variance-maximizing objective in PCA. Assume given a data matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, its covariance matrix $\mathbf{\Sigma} = \mathbf{X}^\top\mathbf{X} \in \mathbb{R}^{n\times n}$, and a single-layer AE projection matrix $\mathbf{W} \in \mathbb{R}^{d\times m}$ with parameters randomly initialized from the continuous uniform distribution $\mathcal{U}(0, 1)$, the reconstruction objective is:

$$\begin{aligned}
\frac{1}{n}\|\mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{W}^\top\|^2 &= \frac{1}{n}\mathrm{tr}((\mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{W}^\top) \\
&\quad \cdot (\mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{W}^\top)^\top) \\
&= \frac{1}{n}\mathrm{tr}((\mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{W}^\top) \\
&\quad \cdot (\mathbf{X}^\top - \mathbf{W}\mathbf{W}^\top\mathbf{X}^\top)) \\
&= \frac{1}{n}[\mathrm{tr}(\mathbf{X}\mathbf{X}^\top) - \mathrm{tr}(\mathbf{X}\mathbf{W}\mathbf{W}^\top\mathbf{X}^\top) \\
&\quad - \mathrm{tr}(\mathbf{X}\mathbf{W}\mathbf{W}^\top\mathbf{X}^\top) \\
&\quad + \mathrm{tr}(\mathbf{X}\mathbf{W}\mathbf{W}^\top\mathbf{W}\mathbf{W}^\top\mathbf{X}^\top)] \\
&= \frac{1}{n}[c_1 - 2\cdot\mathrm{tr}(\mathbf{X}\mathbf{W}\mathbf{W}^\top\mathbf{X}^\top) \\
&\quad + \mathrm{tr}(\hat{\mathbf{X}}\hat{\mathbf{X}}^\top)] \\
&= \frac{1}{n}[c_1 - 2\cdot\mathrm{tr}(\mathbf{X}\mathbf{W}\mathbf{W}^\top\mathbf{X}^\top) + c_2] \\
&= c_3 - c_4\cdot\mathrm{tr}(\mathbf{W}^\top\mathbf{X}^\top\mathbf{X}\mathbf{W}) \\
&= c_3 - c_4\cdot\mathrm{tr}(\mathbf{W}^\top\mathbf{\Sigma}\mathbf{W})
\end{aligned}$$

Notice that c_1, c_2, c_3, c_4 are non-negative scalar constants that do not influence the overall optimization trajectory. Hence, alternatively, the optimal AE projection also maximizes the sample variance $\mathrm{tr}(\mathbf{W}^\top\mathbf{\Sigma}\mathbf{W})$, achieving an identical end goal of PCA transform. Specifically, according to PCA, variance maximization is realized by constructing the projection $\mathbf{W}$ to contain the set of orthonormal eigenvectors of $\mathbf{\Sigma}$ that gives the largest eigenvalues (Hotelling, 1933). That is, there is an orthogonality constraint on $\mathbf{W}$. Minimizing the MSE reconstruction also results in an orthogonal $\mathbf{W}$:

$$\frac{1}{M}\|\mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{W}^\top\|^2 = 0 \Rightarrow \mathbf{W}\mathbf{W}^\top = \mathbf{I}$$

Therefore, the optimal AE projection $\mathbf{W}$ is also capturing a set of variance-maximizing orthogonal vectors. Note that the AE optimized $\mathbf{W}$ is theoretically equivalent to the eigendecomposition of $\mathbf{\Sigma}$ if and only if the reconstruction loss is 0. Therefore, in practice, the AE is, at best, an approximate solution to variance maximization.

A.2. Variance-based Sorting Procedure

Following the discussion in A.1, assuming a perfect optimization, the linear one-layer AE behaves similarly to PCA, and there is an equivalence relation between their respective objective functions. Notice that in PCA, the eigenvalue of the covariance matrix $\mathbf{\Sigma}$ signifies the intensity of data variation along the direction of its corresponding eigenvector, which is essentially a column entry of the transformation matrix. Then intuitively, given an optimized AE projection $\mathbf{W}$, we can examine, for each column of $\mathbf{W}$, its representativeness (*i.e.,* data variance) of the data covariance with a scalar estimate (*i.e.,* an eigenvalue-like scoring). Inspired by the properties of eigendecomposition, we can approximate these estimates by measuring the distance of $\mathbf{W}$ w.r.t to the product of linearly transforming $\mathbf{W}$ through $\mathbf{\Sigma}$ by a scaling factor of $\boldsymbol{\lambda}$. More specifically, we want to solve for $\boldsymbol{\lambda}$ such that $\mathbf{\Sigma}\boldsymbol{w} = \boldsymbol{\lambda}\boldsymbol{w}$ for every column vector $\boldsymbol{w} \in \mathbf{W}$. Under the PCA perspective, $\boldsymbol{\lambda}$ contains the variance estimate for each column-wise individual projection of $\mathbf{W}$. To this end, we detail the sorting procedure in Algorithm 2.

Algorithm 2 Overview procedure for variance-based sorting

Input: Original feature matrix $\mathbf{X} \in \mathbb{R}^{M\times M}$; AE optimized projection matrix $\mathbf{W} \in \mathbb{R}^{M\times D}$

Initialize: Scalar vector $\boldsymbol{\lambda} \in \mathbb{R}^D$; Small positive float ϵ

Output: Sorted AE projection matrix $\tilde{\mathbf{W}}$

1: Normalize the feature matrix: $\mathbf{X_n} \leftarrow \mathbf{X}/\|\mathbf{X}\|$
2: Compute data covariance matrix: $\mathbf{\Sigma} \leftarrow \mathbf{X_n}^\top\mathbf{X_n}$
3: Solve for $\boldsymbol{\lambda}$ such that $|\mathbf{\Sigma}\mathbf{W} - \mathbf{W}\odot\mathrm{diag}(\boldsymbol{\lambda})| \leqslant \epsilon$
4: Sort column vectors $\boldsymbol{w} \in \mathbf{W}$ according to (sorted) decreasing order of $\boldsymbol{\lambda}$ to obtain $\tilde{\mathbf{W}}$

Appendix B. Dataset Details

- **Parkinson's Progression Markers Initiative (PPMI)**: We pre-train the model on large-scale real-life Parkinsons Progression Markers Initiative

(PPMI) data of 718 subjects, where 569 subjects are Parkinson's Disease (PD) patients and the rest 149 are Healthy Control (HC) ones. Eddy-current and head motion correction are performed using FSL[3] and the brain networks are extracted using the same tool. The EPI-induced susceptibility artifacts correction is handled using Advanced Normalization Tools (ANT)[4]. In the meantime, 84 ROIs are parcellated from T1-weighted structural MRI using Freesurfer[5]. The brain networks are constructed using three whole brain tractography algorithms namely the Probabilistic Index of Connectivity (PICo), Hough voting (Hough), and FSL. Each resulted network for each subject is 84×84. Each brain network is normalized by the maximum value to avoid computation bias for the later feature extraction and evaluation, since matrices derived from different tractography algorithms differ in scales and ranges.

- **Bipolar Disorders (BP)**: This local dataset is composed of the resting-state fMRI and DTI image data of 52 Bipolar I subjects who are in euthymia and 45 Healthy Controls (HCs) with matched age and gender (Cao et al., 2015; Ma et al., 2017). The fMRI data was acquired on a 3T Siemens Trio scanner using a T2* echo planar imaging (EPI) gradient-echo pulse sequence with integrated parallel acquisition technique (IPAT) and DTI data were acquired on a Siemens 3T Trio scanner. The brain networks are constructed using the CONN[6] toolbox. We performed the normalization and smoothing after first realigning and co-registering the raw EPI pictures. After that, the signal was regressed to remove the confounding effects of the motion artifact, white matter, and CSF. The 82 cortical and subcortical gray matter regions produced by Freesurfer were identified, and pairwise signal correlations were used to build the brain networks.
- **Human Immunodeficiency Virus Infection (HIV)**: This local dataset involves fMRI and DTI brain networks for 70 subjects, with 35 of them early HIV patients and the other 35 Healthy Controls (HCs). These two groups of subjects do not differ in demographic distributions such as age and biological sex. The preprocessings for fMRI including brain extraction, slice timing correction and realignment are managed with the DPARSF[7] toolbox, while the preprocessings for DTI such as distortion correction are finished with the help of FSL toolbox. Finally, brain networks with 90 regions of interest are constructed based on the automated anatomical labeling (AAL) (Tzourio-Mazoyer et al., 2002).

3. https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/
4. http://stnava.github.io/ANTs/
5. https://surfer.nmr.mgh.harvard.edu/
6. http://www.nitrc.org/projects/conn/

Appendix C. Hyperparameter Setting

GNN Setup. The GCN encoder is composed of 4 graph convolution layers with hidden dimensions of 32, 16, 16, and 8. Similarly, the GAT encoder is built from 4 graph attention layers with hidden dimensions of 32, 16, 16, and 8. Regarding GIN, which is slightly different, the encoder consists of 4 MLP layers with each MLP containing 2 linear layers with a unifying hidden dimension of 8.

Pre-training Pipeline Setup. For two-level node contrastive sampling, we set $k = 2$ as the radius regarding k-hop neighborhood sampling for $\mathbf{S_1}$ and $\mathbf{S_4}$. To enable efficient computation on multi-graph MI evaluation, we resort to mini-batching and we set a default batch size of 32. In addition, we leverage the popular Adam (Kingma and Ba, 2015) optimizer with the learning rate set to 0.002 as well as the cosine annealing scheduler (Loshchilov and Hutter, 2017) to facilitate GNN training. In general, a complete pre-training cycle takes 400 epochs with an active deployment of early stopping.

Atlas Mapping Regularizer Setup. Following the discussion in section 3.3, the total running loss of the AE projection is given as:

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \alpha\mathcal{L}_{\text{loc}} + \beta\mathcal{L}_{\text{com}} + \gamma\mathcal{L}_{\text{KL}}, \tag{7}$$

in particular, we set $\alpha, \beta = 0.8$ and $\gamma = 0.01$. The one-layer AE encoder transforms the feature signals from all given datasets into a universally projected dimension of 32. For the details of locality-preserving regularizer (*i.e.,* $\mathcal{L}_{\text{loc}}$), the transition matrix $\mathbf{T}$ is built from the 5-nearest-neighbor graph from the 3D coordinates of each atlas templates. For the sparsity-oriented regularizer (*i.e.,* $\mathcal{L}_{\text{KL}}$), the target sparsity value ρ is set to $1e^{-5}$. The overall optimization process, which is similar to model pre-training, takes a total of 100 epochs with a learning rate of 0.02.

Downstream Evaluation Setup. For each target evaluation, the fine-tuning process features a 5-fold cross-validation, which approximately splits the

7. http://rfmri.org/DPARSF/

Table 4: Disease prediction performance of our framework using GAT and GIN. The best performer is highlighted in bold.

Method	BP-fMRI		BP-DTI		HIV-fMRI		HIV-DTI	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
Ours w/ GCN	**68.84**$_{\pm 8.26}$	68.45$_{\pm 8.96}$	**66.57**$_{\pm 7.67}$	**68.31**$_{\pm 9.39}$	**77.80**$_{\pm 9.76}$	77.22$_{\pm 8.74}$	**67.51**$_{\pm 8.67}$	**67.74**$_{\pm 8.59}$
Ours w/ GAT	66.96$_{\pm 9.71}$	**69.68**$_{\pm 9.61}$	64.23$_{\pm 10.47}$	63.76$_{\pm 10.49}$	74.93$_{\pm 10.35}$	75.78$_{\pm 11.12}$	65.84$_{\pm 9.74}$	66.51$_{\pm 12.07}$
Ours w/ GIN	66.30$_{\pm 8.77}$	68.92$_{\pm 9.37}$	64.48$_{\pm 9.83}$	66.44$_{\pm 8.58}$	75.96$_{\pm 9.56}$	**77.63**$_{\pm 10.10}$	67.36$_{\pm 9.26}$	65.95$_{\pm 11.76}$

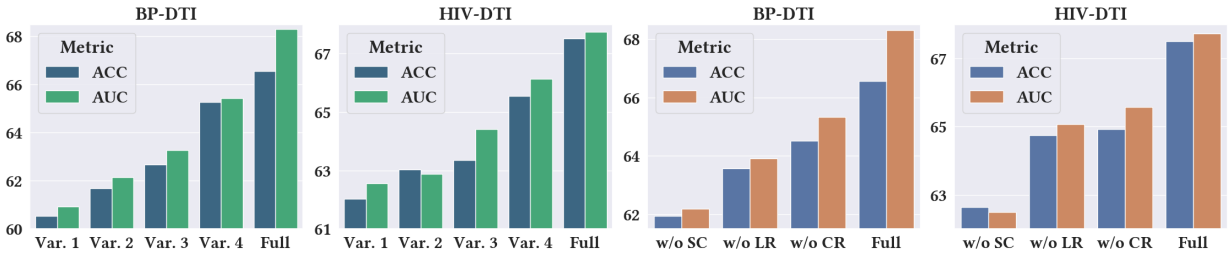


Figure 7: Additional ablation comparisons on DTI views. The left two subfigures refer to contrastive sampling considerations and the right two subfigures refer to atlas mapping regularizers. The y-axis refers to the numeric values of evaluated metrics (in %). We benchmark our results on the DTI modality of the BP and HIV dataset in this Appendix.

dataset into 70% training, 10% validation, and 20% testing. To prevent model over-fitting, we implement a L_2 penalty with a coefficient of $1e^{-4}$. Overall, the model fine-tuning process, which is nearly identical to the other two training procedures, takes a total of 200 epochs with a learning rate of 0.001 and a cosine annealing scheduler.

Appendix D. Additional Experiment

D.1. Performance with GAT and GIN

Table 4 reports the downstream performance of our full framework using GAT and GIN as backbone encoders. In general, the two encoders deliver inferior performance compared to GCN, which suggests that complex GNN convolutions (*e.g.,* GAT and GIN) might not be as effective as they seem when learning on brain network datasets.

D.2. Additional Ablation Studies on DTI

Figure 7 presents our ablation studies on the DTI view following the same setup as discussed in Section 4.2. We draw similar conclusions from the DTI-based analysis where each constituent component of our two-level sampling consideration as well as the atlas mapping mechanism has proven positive contribution and significance towards the overall performance and robustness.