# Disentangled and Proportional Representation Learning for Multi-View Brain Connectomes

Yanfu Zhang[1], Liang Zhan[1], Shandong Wu[2], Paul Thompson[3], and Heng Huang[1,4]

[1] Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA 15260, USA
[2] Department of Radiology, University of Pittsburgh, Pittsburgh, PA 15260, USA
[3] Imaging Genetics Center, Institute for Neuroimaging and Informatics, University of Southern California, Los Angeles, CA 90032, USA
[4] JD Finance America Corporation, Mountain View, CA 94043, USA

**Abstract.** Diffusion MRI-derived brain structural connectomes or brain networks are widely used in the brain research. However, constructing brain networks is highly dependent on various tractography algorithms, which leads to difficulties in deciding the optimal view concerning the downstream analysis. In this paper, we propose to learn a unified representation from multi-view brain networks. Particularly, we expect the learned representations to convey the information from different views fairly and in a disentangled sense. We achieve the disentanglement via an approach using unsupervised variational graph auto-encoders. We achieve the view-wise fairness, *i.e.* proportionality, via an alternative training routine. More specifically, we construct an analogy between training the deep network and the network flow problem. Based on the analogy, the fair representations learning is attained via a network scheduling algorithm aware of proportionality. The experimental results demonstrate that the learned representations fit various downstream tasks well. They also show that the proposed approach effectively preserves the proportionality.

**Keywords:** Brain Connectome · Alzheimer's Disease · Multi-view · Prediction.

## 1 Introduction

Human brain connectomes [6] are models of complex brain networks and can be derived from diverse experimental modalities and tractography algorithms. Large-scale brains connections convey important insights for understanding the underlying yet largely unknown mechanisms of many mental disorders [11,15,26,7]. Nevertheless, the apparent characteristics of brain networks are profoundly influenced by the tractography algorithms. The designs of tractography algorithms, including tensor-based deterministic algorithms [2], probabilistic approaches [18], random forest [17] and Deep Neural Network (DNN) [20], and regularized methods guided by biologically plausible fascicle structures [3], are inspired by specific

experimental questions [5], *e.g.*, different tractography algorithms are used for predicting or classifying neurodegenerative or neurodevelopmental conditions based on various brain abnormalities. For example, the selection and accuracy of the extracted fibers are different for different tractography algorithms, and the relevance of the extracted fiber bundles depend on the different tasks and questions being addressed. Therefore, it is elusive to decide a universally optimal modality of brain networks and associated processing pipeline for distinct diagnostic tasks [5,23].

Multi-view methods can leverage the available information from diverse tractography algorithms simultaneously, and tentative studies have demonstrated that multi-modal brain networks can provide complementary viewpoints for the classification tasks, *e.g.*, multi-view graph convolutional network [25] is found to have state-of-the-art performance in classifying Parkinson's disease (PD) status. However, previous multi-view methods have two restrictions regarding general prediction tasks of neurodegenerative conditions. First, many methods are designed for some specific tasks. If one want to tailor these methods to other tasks, it is necessary to carefully tune the hyperparameters. Second, though some methods learn representations from multi-view brain networks, the learning is guided by some predefined prediction tasks, which may introduce bias to overemphasize a particular modality. As such, the learned embeddings cannot represent multi-modal brain networks comprehensively, and their application to the related analysis in a broader scope is potentially constrained.

To address these problem, we propose to learn unified representations from multi-modal brain networks via unsupervised learning techniques. To extend the generalization ability of the learned representations to different downstream analysis, the representations shall be of *disentanglement* and *proportionality* concerning different modalities. Here, disentanglement refers to the representations encoding salient attributes of data explicitly, which can help the analysis of the prediction tasks and the modalities. Proportionality refers to a balanced contribution to the representations of each modality, which avoids the potential bias on specific modalities. In other words, in our approach the learned representations can fairly convey the information from different modalities and can be exploited by various downstream analysis. More specifically, in this paper we propose a multi-view graph auto-encoder to learn the disentangled graph embeddings from brain networks. We formulate the proportionality-awareness in multi-view representation learning as a network scheduling problem via an analogy between training deep networks and the graph flow problems. The experimental results demonstrate the effectiveness of the proposed method.

## 2 Methodology

The proposed method is illustrated in Fig. 1. For each view, a Variational Graph Auto-encoder (VGAE) [10] is exploited. Let $G^{(v)}$ denote the brain networks of the $v^{th}$ view, $f^{(v)}$ and $g^{(v)}$ the corresponding encoder and decoder, $[\mu^{(v)}|\sigma^{(v)}] = f(G^{(v)})$ is the estimated mean and variance of the encoder. The

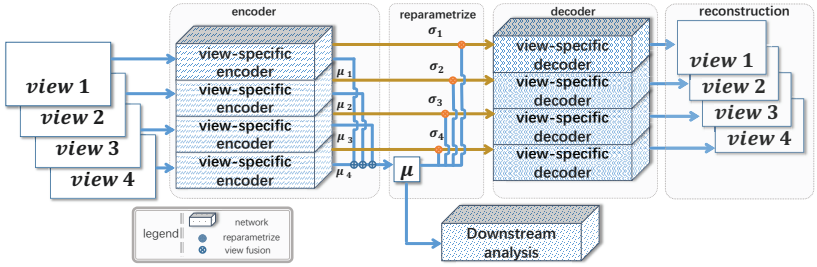


Fig. 1: The structure of the proposed method. Each view uses an independent VGAE to learn a unified $\mu$, while the $\sigma$ is different.

unified representations are computed by max-out the stacked $\mu^{(v)}$ by the position, which can be denoted as $\mu = maxpool1d([\mu^{(v)}])$. The reparameterization for the $v^{th}$ view is then computed using $\mu$ and $\sigma^{(v)}$. $\mu \in \mathbb{R}^k$ is also used as the embeddings. According to the structure of VGAE, $\sigma^{(v)} \in \mathbb{R}^k$. Besides the view-wise VGAE loss, we push $\mu$ and $\mu^{(v)}$ to be close so that the learned embeddings for different views are consistent. The disentanglement of the representations is acquired via introduce the $\beta$-VAE loss [8]. Disentangled representations are compact and interpretable [4]. The objective for our multi-view GVAE is:

$$\mathcal{L} = \sum_{v \in \mathcal{V}} \mathbb{B}\left(log\left(P(\tilde{G}^{(v)})\right)\right) + \beta KL\left(P(z^{(v)})|\mathcal{N}(0,1)\right)\right) + \lambda(\mu^{(v)} - \mu)^2 \qquad (1)$$

here the first term is the reconstruction loss, the second is the Kullback-Leibler divergence, and the last is the multi-view consistency.

As aforementioned, the representations shall also be fair to different views. In the above auto-encoder framework, the decoder is used for evaluating the vividness of the learned representations. However, for multi-view data, the reconstruction for different views is not necessarily equally accurate. When the imbalance occurs, some views are less included in the learned representations. To address this problem, we consider to learn fair representations regarding different views, which indicates *the view-wise loss in* (1) *is close to each other*. Such fairness, referred to as *proportionality*, can be achieved via an alternative training routine of the above model. We will formulate an analog between flow network problem and the training of multi-view model in the following. Based on the formulation, we design a scheduling algorithm to satisfy the proportionality requirement.

**Training Multi-view Network: a Flow Network Perspective** Directed Acyclic Graph (DAG) is an important tool in graphical models [9]. It is also exploited to express network structures by many popular deep learning frameworks [19,1]. Inspired by this idea, we make an analogy between training the deep network and the flow network problems.

In Fig. 2, we illustrate an example for multi-view learning. To simplify the elaboration, we consider a structure taking two views $s_1$ and $s_2$, as inputs. The network consists of four sub-networks, each corresponding to one edge in the DAG. $v_0$ is a fused hidden representation, and $t_1$ is the prediction. For multiple inputs, $\oplus$ denotes the fusion operation for the outputs of multiple sub-networks,

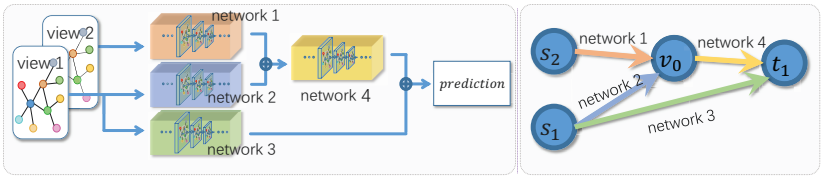


Fig. 2: Left: a simple DNN. Right: the corresponding DAG. Each edge represents a network, and each node denote an intermediate representation.

and it can either a weighted summation or concatenation. Consider a network trained after $t$ steps using gradient based method. In the $t+1$ step, we can define the flow $d_{i,j}$ from predecessor $i$ to successor $j$ as $d_{i,j} = \triangle\mathcal{L}\left(f_j^{(t+1)}(h_i^{(t+1)}, \mathcal{H}_{j\backslash i}^{(t)})\right)$, here $\mathcal{L}$ is an objective defined on the targets, and $\triangle\mathcal{L}$ denotes the loss difference between step $t+1$ and $t$. Let $\mathcal{P}_{ij}$ represent the set of all paths from sources to targets containing $e_{i,j}$. $f_j^{(t+1)}$ refers to the network to compute the final outputs with all paths in $\mathcal{P}_{ij}$ updated. $\mathcal{P}_{ij}$ can be defined on the node $i$ and a set $\mathcal{H}_{j\backslash i}$. Here $\mathcal{H}_{j\backslash i}$ denotes any cut set containing node $j$ that separate sources and targets, and $\mathcal{H}_{j\backslash i}$ does not include any node in $\mathcal{P}_{ij}$ except $j$.

Our definition satisfies the flow conservation, which states that if a node is neither a source or a target, its net flow shall be 0. For a node $j$ with multiple incoming flow, the fusion operation is defined as $\boldsymbol{h}_j = \sum_{i\in\mathcal{P}_j} \boldsymbol{P}_{ij}\boldsymbol{W}_{ij} f_{ij}(\boldsymbol{h}_i)$, here $\mathcal{P}_j$ is the predecessor set of node $j$, $f_{ij}$ is the sub-network between node $i$ and $j$. For different fusion oeprations, $\boldsymbol{P}_i$ and $\boldsymbol{W}_i$ take different forms. For example, when both $\boldsymbol{P}_i$ and $\boldsymbol{W}_i$ are the identity matrices, the fusion is by summation; if $\boldsymbol{W}_i$ is the augmented matrix $(\boldsymbol{I}_i|\boldsymbol{0})$, fusion by concatenation is feasible by setting $\boldsymbol{P}_i$ as the corresponding permutation matrix. For a node with multiple outgoing flow, the output is equally distributed. We abuse the notation $\mathcal{H}_j \equiv \mathcal{H}_{j\backslash i} \cup \{i\}$. Consider a fixed given cut $\mathcal{H}_j$ for node $j$, we can induce two additional cuts: $\mathcal{H}_{\mathcal{P}_j}$, which excludes $j$ and include all its predecessors; and $\mathcal{H}_{\mathcal{S}_j}$, which excludes $j$ and include all its successors. Under the updating rule of backward propagation, the incoming flow with respect to node $j$ is,

$$\sum_{i\in\mathcal{P}_j} d_{i,j} \approx \frac{\partial\mathcal{L}}{\partial f_j}\sum_{i\in\mathcal{P}_j} \boldsymbol{P}_{ij}\boldsymbol{W}_{ij}\frac{\partial f_j}{\partial \boldsymbol{h}_i}d\boldsymbol{h}_i = \frac{\partial\mathcal{L}}{\partial f_j}\frac{\partial f_j}{\partial \boldsymbol{h}_j}d\boldsymbol{h}_j, \tag{2}$$

the above equation follows because the partial differential is 0 except $d\boldsymbol{h}_i$ and $d\boldsymbol{h}_j$ term. Similarly, the outgoing flow is,

$$\sum_{k\in\mathcal{S}_j} d_{j,k} \approx \sum_{k\in\mathcal{S}_j} \frac{\partial\mathcal{L}}{\partial f_k}\boldsymbol{P}_{jk}\boldsymbol{W}_{jk}\frac{\partial f_k}{\partial \boldsymbol{h}_j}d\boldsymbol{h}_j = \frac{\partial\mathcal{L}}{\partial f_j}\frac{\partial f_j}{\partial \boldsymbol{h}_j}d\boldsymbol{h}_j, \tag{3}$$

(2) and (3) are bridged by the change in $h_j$, which ensures the net flow to be 0.

If we extend the above analogy to the accumulative case, the flow is defined to be the loss decrease with respect to the particular structure represented by $i \to j$. Noteworthy, it is not the pure contribution of $i \to j$. Rather, it is more of the quantification of the total loss decrease of the particular structure, as

Round-Robin

**Input:** $v$ views, max epoch $e$
**Output:** model $f$
Initialize $f$.
**repeat**
  **for** $i \leftarrow 1$ **to** $v$ **do**
    Optimize (1) *w.r.t.* view $j$.
**until** *max epoch*;

Proportionality

**Input:** $v$ views, max epoch $e$
**Output:** model $f$
Initialize $f$.
**repeat**
  Compute priority *w.r.t.* (6)
  Optimize (1) *w.r.t.* view $j$ with the highest priority.
**until** *max epoch*;

the definition considers both the upstream and downstream computation of the entire network. The empirical loss is related to the generalization bound of the learned representations concerning downstream tasks. As such, the accumulated flow can be interpreted as the amount of information learned from each view informally. Based on this analogy, we define that the proportionality is achieved if the view-wise flow, *i.e.* the accumulated $\sum_{k\in\mathcal{S}_j} d_{j,k}$ for some view $j$, is balanced.

**Alternative Training Routine with Proportionality Awareness** Conventionally, the proportionality concerning different views can be written as a constrained optimization problem, and a standard training routine is based on SGD. From the flow perspective, the proportional training can be interpreted as multiple views competing for the updating resources in the backward propagation, which is a network scheduling algorithm. More specifically, during the training, the accumulated flow is continuously updated, which reflexes the dynamic of loss decrease and the generalization ability. A proportional representation is then equivalent to a balanced flow avoiding the overload of some specific path.

In detail, we define the total flow as the loss decrease. When the learning rate is small enough, the summation of view-wise SGD update is equivalent to a *round-robin* update with respect to each view. Here, the objective associated with each view is optimized in a predefined turn. To avoid a specific view taking up too much updating resources, we can maximize the total flow of the network while allowing the minimal level of service for all views via introducing a competing mechanism for each view to occupy the update based on the estimated flow. We refer to this method as *proportionality*. The updating priority of each view is based on the current loss decrease and the historical cumulative loss decrease. Assume the loss decrease of view $i$ at update $t$ can be foreseen as $r_{i,t}$. The throughput of view $i$ is defined as historical cumulative loss decrease at step $t$:

$$\theta_{i,t} = \theta_{i,0} + \sum_{l=1}^{t} \frac{r_{i,l} I_{i,l}}{t} = \frac{n-1}{n}\theta_{i,t-1} + \frac{1}{n} r_{i,t-1} I_{i,t-1}, \tag{4}$$

where $I_{i,l}$ is an indicator. $I_{i,l} = 1$ if the $l^{th}$ update is conducted on view $i$, and 0 otherwise. Based on (4), the priority $p_{i,t}$ for view $i$ can be defined, and the $t+1$ update is then applied to the view with the highest priority:

$$\arg\max_{i\in\mathcal{V}} \{p_{i,t}\}, \qquad p_{i,t} = \frac{r_{i,t+1}}{\epsilon + \theta_{i,t}} \tag{5}$$

where $\epsilon$ is a small positive number for computational stability. Notably, the above algorithms is not immediately applicable to our formulation, as that $r_{i,t}$ is not pre-assigned as in standard proportionally fairness algorithms. Instead, the values are only known after the update is finished. Thus, we propose a compensation update method: at the beginning, we use one round robin update and compute initial $r_{i,0}$. In the following steps we use proportionally fairness algorithm, but computing the priority using the loss decrease from the last applied update:

$$\arg\max_{i \le v} \left\{ \frac{r_{i,t_i}}{d_i + \theta_{i,t}} \right\}, \quad t_i = \max l, \quad s.t. \quad l \le t, \quad I_{i,l} = 1, \tag{6}$$

The proportionality and convergence of our scheduling algorithm are guaranteed under some weak conditions, and the analysis can be found in [13].

## 3 Experimental Results

In this experiments we use three datasets, including the data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) and National Alzheimer's Coordinating Center (NACC), and the Parkinson Progression Marker Initiative (PPMI). The preprocessed ADNI brain networks [22] include 51 healthy controls (HC) (mean age=$69.69 \pm 10.27$, 29 males), 112 people with Mild Cognitive Impairment (MCI) (mean age=$71.68 \pm 9.89$, 41 males) and 39 individuals with AD (mean age=$75.56 \pm 8.99$, 14 males). The similarly preprocessed NACC brain networks [21] include 329 HCs (mean age=$60.96 \pm 8.96$, 107 males), 57 with MCI (mean age=$73.60 \pm 7.93$, 38 males), and 54 AD patients (mean age = $72.02 \pm 10.41$, 32 males). The similarly preprocessed PPMI brain networks [27,28] includes 145 HC (mean $age = 66.70 \pm 10.95$, 96 males) and 474 subjects with PD (mean age=$67.33 \pm 9.33$, 318 males). Nine different views are reconstructed using T-FACT, T-RK2, T-TL, T-SL, O-FACT and O-RK2, Probt, Hough, and PICo (Please refer to [24] for more details on the brain network reconstruction). We use a modified network structure based on graph variational auto-encoder. The view-wise graph is the averaged brain connectome, and the node features are the corresponding row for each brain connectome. We set $\beta = 4$ recommended by $\beta$-VAE [8]. The performance is not sensitive to $\lambda$, and we set it to 0.001. In the encoder, we use three graph convolutional layers for $\mu$ and $\sigma$ respectively. The first two layers are shared, both with 64 hidden units. The embedding length is 32. The encoder are limited in layers due to the potential over-smoothing for graph convolutional layers. Our model is trained 100 epochs using ADAM with batch size 32 and learning rate 0.0001.

**Evaluating the Proposed Method in Down-streaming Analysis:** We compare our approach with related baselines on several classification and regression tasks. The ablation study is also included.

Table 1 summarizes the classification results. For ADNI and NACC, we predict the HC and AD. For PPMI we predict HC and PD. For multi-view predictions, we include principal component analysis (PCA), multi-view non-negative matrix factorization (MVNMF) [14], co-regularized spectral clustering

Table 1: The comparison on classification tasks.

Sparse Logistic Regression

| | | ADNI | NACC | PPMI |
|---|---|---|---|---|
| Single View | FSL | $0.7786 \pm 0.0976$ | $0.7669 \pm 0.0799$ | $\mathbf{0.6597 \pm 0.0584}$ |
| | PICo | $0.7615 \pm 0.1408$ | $0.7119 \pm 0.1103$ | $0.6065 \pm 0.0486$ |
| | T-FACT | $0.7451 \pm 0.0379$ | $0.6581 \pm 0.0411$ | $0.5850 \pm 0.0433$ |
| | O-FACT | $0.7278 \pm 0.1066$ | $0.7094 \pm 0.0866$ | $0.5921 \pm 0.0353$ |
| | ODF-Rk2 | $0.7568 \pm 0.0821$ | $0.6890 \pm 0.0366$ | $0.5942 \pm 0.0331$ |
| | T-RK2 | $0.7276 \pm 0.0797$ | $0.7281 \pm 0.0674$ | $0.5921 \pm 0.0353$ |
| | T-SL | $0.7402 \pm 0.1371$ | $0.6582 \pm 0.0785$ | $0.5884 \pm 0.0389$ |
| | T-TL | $0.6875 \pm 0.0682$ | $0.7358 \pm 0.0799$ | $0.5851 \pm 0.0423$ |
| | Hough | $0.7559 \pm 0.0780$ | $0.7271 \pm 0.0549$ | $0.5536 \pm 0.0391$ |
| Multi View | all views | $0.7966 \pm 0.0904$ | $0.7301 \pm 0.1325$ | $0.5716 \pm 0.0378$ |
| | MVNMF | $0.8149 \pm 0.0550$ | $0.7685 \pm 0.0958$ | $0.6104 \pm 0.0332$ |
| | MVSC | $0.8203 \pm 0.0791$ | $0.7595 \pm 0.1013$ | $0.6205 \pm 0.0373$ |
| | DMGCN | $0.8058 \pm 0.1006$ | $0.7557 \pm 0.0898$ | $0.6141 \pm 0.0707$ |
| | Proposed-I | $0.8074 \pm 0.0493$ | $0.7491 \pm 0.0897$ | $0.6122 \pm 0.0442$ |
| | Proposed-II | $0.8185 \pm 0.0770$ | $0.7549 \pm 0.0790$ | $0.6240 \pm 0.0234$ |
| | proposed* | $\mathbf{0.8278 \pm 0.1537}$ | $\mathbf{0.8090 \pm 0.1472}$ | $0.6250 \pm 0.0472$ |

Random Forest

| | | ADNI | NACC | PPMI |
|---|---|---|---|---|
| Single View | FSL | $0.8124 \pm 0.0455$ | $0.3737 \pm 0.7065$ | $0.5753 \pm 0.0255$ |
| | PICo | $0.7838 \pm 0.1067$ | $0.1588 \pm 0.9463$ | $0.5475 \pm 0.0244$ |
| | T-FACT | $0.8383 \pm 0.0483$ | $0.7029 \pm 0.1184$ | $0.5654 \pm 0.0331$ |
| | O-fact | $0.7817 \pm 0.1512$ | $0.3789 \pm 0.6903$ | $0.5478 \pm 0.0228$ |
| | O-RK2 | $0.7617 \pm 0.1087$ | $0.7879 \pm 0.1333$ | $0.5566 \pm 0.0382$ |
| | T-RK2 | $0.7764 \pm 0.1275$ | $0.7029 \pm 0.1333$ | $0.5486 \pm 0.0361$ |
| | T-SL | $0.8148 \pm 0.0587$ | $0.7235 \pm 0.1163$ | $0.5386 \pm 0.0347$ |
| | T-TL | $0.7695 \pm 0.0862$ | $0.7009 \pm 0.1164$ | $0.5411 \pm 0.0410$ |
| | Hough | $0.8368 \pm 0.0671$ | $0.7011 \pm 0.1797$ | $0.5276 \pm 0.0344$ |
| Multi View | all views | $0.8560 \pm 0.0574$ | $0.7615 \pm 0.1053$ | $0.5743 \pm 0.0464$ |
| | MVNMF | $0.8826 \pm 0.0830$ | $0.8317 \pm 0.1561$ | $0.5659 \pm 0.0528$ |
| | MVSC | $0.8827 \pm 0.0457$ | $0.7997 \pm 0.1435$ | $0.5753 \pm 0.0348$ |
| | DMGCN | $0.8862 \pm 0.0503$ | $0.8307 \pm 0.1493$ | $0.5683 \pm 0.0323$ |
| | Proposed-I | $0.8578 \pm 0.0516$ | $0.7919 \pm 0.0725$ | $0.5590 \pm 0.0250$ |
| | Proposed-II | $0.8678 \pm 0.0573$ | $0.8327 \pm 0.0988$ | $0.5699 \pm 0.0382$ |
| | Proposed* | $\mathbf{0.8946 \pm 0.0510}$ | $\mathbf{0.8359 \pm 0.1321}$ | $\mathbf{0.5814 \pm 0.0274}$ |

(MVSC) [12] and Deep Metric Graph Convolutional Network (DMGCN) [11]. We use the aforementioned methods to learn the representations, and then exploit two off-the-shelves methods, sparse logistic regression, and random forest to make the final prediction. We report AUC on 5-fold cross-validation. To make the comparison self-contained, single view results are also included. For the ablation study, in *propose-I* neither disentanglement nor proportionality is considered, and

Table 2: The comparison on regression tasks.

| | | TD | UPSIT | MoCA |
|---|---|---|---|---|
| Single View | FSL | $0.0749 \pm 0.0167$ | $0.0794 \pm 0.0054$ | $0.0381 \pm 0.0033$ |
| | PICo | $0.0714 \pm 0.0078$ | $0.0983 \pm 0.0101$ | $0.0394 \pm 0.0027$ |
| | T-FACT | $0.0404 \pm 0.0037$ | $0.0545 \pm 0.0043$ | $0.0210 \pm 0.0021$ |
| | O-FACT | $0.0410 \pm 0.0018$ | $0.0508 \pm 0.0066$ | $0.0208 \pm 0.0036$ |
| | O-RK2 | $0.0428 \pm 0.0079$ | $0.0503 \pm 0.0015$ | $0.0208 \pm 0.0027$ |
| | T-RK2 | $0.0441 \pm 0.0034$ | $0.0500 \pm 0.0051$ | $0.0212 \pm 0.0025$ |
| | T-SL | $0.0427 \pm 0.0012$ | $0.0512 \pm 0.0058$ | $0.0212 \pm 0.0017$ |
| | T-TL | $0.0406 \pm 0.0059$ | $0.0517 \pm 0.0019$ | $0.0210 \pm 0.0027$ |
| | Hough | $0.0434 \pm 0.0036$ | $0.0495 \pm 0.0058$ | $0.0225 \pm 0.0044$ |
| Multi View | all views | $0.0414 \pm 0.0045$ | $0.0524 \pm 0.0034$ | $0.0227 \pm 0.0074$ |
| | MVNMF | $0.0378 \pm 0.0122$ | $0.0507 \pm 0.0041$ | $0.0207 \pm 0.0030$ |
| | MVSC | $0.0355 \pm 0.0047$ | $0.0499 \pm 0.0046$ | $\mathbf{0.0199 \pm 0.0013}$ |
| | DMGCN | $0.0365 \pm 0.0071$ | $0.0487 \pm 0.0085$ | $0.0202 \pm 0.0015$ |
| | Proposed-I | $0.0358 \pm 0.0024$ | $0.0501 \pm 0.0022$ | $0.0209 \pm 0.0019$ |
| | Proposed-II | $0.0361 \pm 0.0035$ | $0.0492 \pm 0.0033$ | $0.0200 \pm 0.0022$ |
| | Proposed* | $\mathbf{0.0351 \pm 0.0059}$ | $\mathbf{0.0484 \pm 0.0044}$ | $\mathbf{0.0199 \pm 0.0022}$ |

in *proposed-II* the disentanglement is considered. The full approach is *proposed**. We omit more single-view ablation study in the experiments because our objective is designed for multi-view data. Of note, integrating multi-view data is also shown to be beneficial for brain network analysis [27]. From the results, we find that the prediction ability of different views with respect to different tasks are complicated, and heavily coupled with the algorithms. Multi-view methods, generally, can improve the prediction ability. However, the advantage of multi-view data is intriguing and needs careful examination. The proposed method have good performances and are robust with respect to different tasks. And at last, the ablation study demonstrates that the performance can be improved through considering disentanglement and proportionality.

Table 2 summarizes the regression results. We use the learned representation to predict several clinical scores, including the Tremor Dominant scores (TD), the University of Pennsylvania Smell Identification Test (UPSIT), and the Montreal Cognitive Assessment Test (MoCA). Mean squared error (MSE) is used as the metric evaluating the prediction. All scores are normalized to $[0, 1]$. The results show that the prediction is more complicated with respect to the particular medical scores and views. Similarly, we can observe the advantage of utilizing multi-view data and the robust and superior prediction abilities of our approach.

**Evaluating the Proportionality during Training:** In this section, we demonstrate the proposed method can achieve proportionality using the proposed training scheduling method. Figure 4 illustrates the training loss of the proposed deep network against epochs, and the shaded area represents the variance regarding different views. From the results, we can observe that the proposed method effectively reduces the variance during training, which indicates the learned representations proportionally represent different modalities of brain networks. The

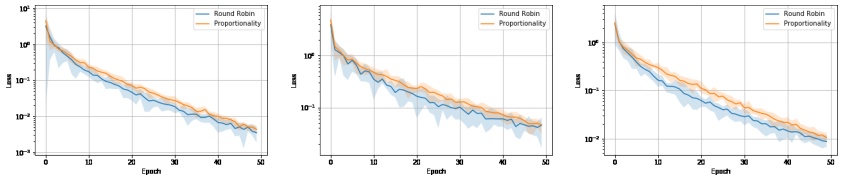


Fig. 4: Left to right: ADNI, NACC, PPMI.

results also show the training routine aware of proportionality converges slightly slower than the standard training routine. However, with moderate epochs their performance difference is negligible.

**Discussions:** There some works applying the fairness principle on brain analysis [16]. Our method is designed for representation learning for multi-view brain connectomes, particularly focusing on the disentangled and proportional property (which is related to algorithmic fairness) for the learned embeddings. Our experimental results demonstrate that the proposed method can be applied to various downstream works. As such, it is of potential to apply our method to broader applications, including generating a refined connectome matrix,

## 4 Conclusion

In this paper, we propose an unsupervised method to learn unified graph embeddings for multi-view brain networks. We design a multi-view graph variational auto-encoder to learn the representations with disentanglement and proportionality. The experimental results demonstrate that the learned representations can be effectively used by various downstream tasks.

## Acknowledgements

This work was partially supported by NSF IIS 1845666, 1852606, 1838627, 1837956, 1956002, 2045848, IIA 2040588, and NIH U01AG068057, R01AG049371, R01AG071243, RF1MH125928.

The NACC database was funded by NIA U01AG016976. The ADNI data were funded by the Alzheimer's Disease Metabolomics Consortium (NIA R01AG046171, RF1AG051550 and 3U01AG024904-09S4). The PPMI data were obtained from the Parkinson's Progression Markers Initiative (PPMI) database.

## References

1. M. Abadi, , et al. Tensorflow: A system for large-scale machine learning. In *OSDI*, pages 265–283, 2016.
2. I. Aganj et al. A hough transform global probabilistic approach to multiple-subject diffusion mri tractography. *Medical image analysis*, 15(4):414–425, 2011.
3. F. Aminmansour et al. Learning macroscopic brain connectomes via group-sparse factorization. In *Advances in Neural Information Processing Systems*, pages 8847–8857, 2019.
4. Y. Bengio et al. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

5. E. Bullmore et al. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature reviews neuroscience*, 10(3):186–198, 2009.
6. E. T. Bullmore and D. S. Bassett. Brain graphs: graphical models of the human brain connectome. *Annual review of clinical psychology*, 7:113–140, 2011.
7. C. Caspell-Garcia et al. Multiple modality biomarker prediction of cognitive impairment in prospectively followed de novo parkinson disease. *PloS one*, 12(5):e0175674, 2017.
8. I. Higgins et al. beta-vae: Learning basic visual concepts with a constrained variational framework. *Iclr*, 2(5):6, 2017.
9. F. Huang and S. Chen. Learning dynamic conditional gaussian graphical models. *IEEE Transactions on Knowledge and Data Engineering*, 30(4):703–716, 2017.
10. T. N. Kipf and M. Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016.
11. S. I. Ktena et al. Distance metric learning using graph convolutional networks: Application to functional brain networks. In *MICCAI*, pages 469–477, 2017.
12. A. Kumar, P. Rai, and H. Daume. Co-regularized multi-view spectral clustering. In *Advances in neural information processing systems*, pages 1413–1421, 2011.
13. H. J. Kushner et al. Convergence of proportional-fair sharing algorithms under general conditions. *IEEE T. on Wireless Communications*, 3(4):1250–1259, 2004.
14. J. Liu, C. Wang, J. Gao, and J. Han. Multi-view clustering via joint nonnegative matrix factorization. In *SIAM International Conference on Data Mining*, pages 252–260. SIAM, 2013.
15. L. Luo, J. Xu, C. Deng, and H. Huang. Robust metric learning on grassmann manifolds with generalization guarantees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4480–4487, 2019.
16. D. Moyer, G. Ver Steeg, C. M. Tax, and P. M. Thompson. Scanner invariant representations for diffusion mri harmonization. *Magnetic resonance in medicine*, 84(4):2174–2189, 2020.
17. P. F. Neher et al. A machine learning based approach to fiber tractography using classifier voting. In *MICAI*, pages 45–52, 2015.
18. G. J. Parker et al. A framework for a streamline-based probabilistic index of connectivity (pico) using a structural interpretation of mri diffusion measurements. *Journal of Magnetic Resonance Imaging*, 18(2):242–254, 2003.
19. A. Paszke et al. Automatic differentiation in pytorch. 2017.
20. P. Poulin et al. Learn to track: Deep learning for tractography. In *MICCAI*, pages 540–547, 2017.
21. Q. Wang, L. Guo, P. M. Thompson, C. R. Jack Jr, H. Dodge, L. Zhan, J. Zhou, A. D. N. Initiative, et al. The added value of diffusion-weighted mri-derived structural connectome in evaluating mild cognitive impairment: A multi-cohort validation. *Journal of Alzheimer's Disease*, 64(1):149–169, 2018.
22. Q. Wang, M. Sun, L. Zhan, P. Thompson, S. Ji, and J. Zhou. Multi-modality disease modeling via collective deep matrix factorization. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1155–1164, 2017.
23. Q. Wang, L. Zhan, P. M. Thompson, H. H. Dodge, and J. Zhou. Discriminative fusion of multiple brain networks for early mild cognitive impairment detection. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 568–572. IEEE, 2016.
24. L. Zhan, J. Zhou, Y. Wang, Y. Jin, N. Jahanshad, G. Prasad, T. M. Nir, C. D. Leonardo, J. Ye, P. M. Thompson, et al. Comparison of nine tractography algorithms

for detecting abnormal structural brain networks in alzheimer's disease. *Frontiers in aging neuroscience*, 7:48, 2015.
25. X. Zhang et al. Multi-view graph convolutional network and its applications on neuroimage analysis for parkinson's disease. In *AMIA Annual Symposium Proceedings*, volume 2018, page 1147, 2018.
26. Y. Zhang and H. Huang. New graph-blind convolutional network for brain connectome data analysis. In *International Conference on Information Processing in Medical Imaging*, pages 669–681. Springer, 2019.
27. Y. Zhang, L. Zhan, W. Cai, P. Thompson, and H. Huang. Integrating heterogeneous brain networks for predicting brain disease conditions. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 214–222. Springer, 2019.
28. Y. Zhang, L. Zhan, P. M. Thompson, and H. Huang. Biological knowledge guided deep neural network for brain genotype-phenotype association study. In *Multimodal Brain Image Analysis and Mathematical Foundations of Computational Anatomy*, pages 84–92. Springer, 2019.