EMORY GRAPH MINING GROUP

Interpretable Graph Neural Networks for Connectome-Based Brain Disorder Analysis

Hejie Cui[1], Wei Dai[1], Yanqiao Zhu[2], Xiaoxiao Li[3], Lifang He[4], and Carl Yang[1]

Presenter: Hejie Cui

Paper *Code*

EMORY UNIVERSITY

UCLA

UBC THE UNIVERSITY OF BRITISH COLUMBIA

LEHIGH UNIVERSITY

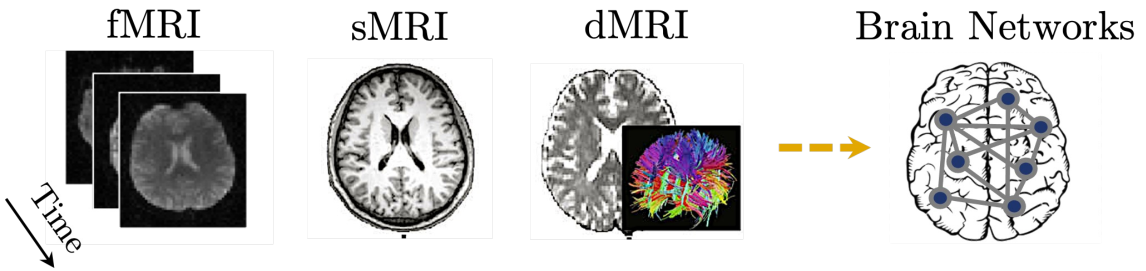
Outline

1. **Introduction & Motivation**
2. **The Proposed Model: IBGNN+**
   - *The backbone model IBGNN*
   - *The globally shared explanation generator*
   - *Enhancing the backbone with the learned explanations*
3. **Experiments**
4. **Interpretation Analysis**
5. **Conclusion & Future Work**

Introduction: Brain Networks & GNNs

Brain Networks

- Human brains lie at the core of neurobiological systems
- Mapping the connections of the human brain as a **network** is one of the most pervasive paradigms in neuroscience
  - *Nodes: anatomic regions*
  - *Edges: connectivities between the regions*
- **Interpretable** models on brain networks are vital



Graph Neural Networks (GNNs)

- GNNs have emerged and proved its power for analyzing graph-structured data.
- Compared with shallow models → universal expressiveness to capture the sophisticated connectome structures
- However, as a family of deep models

Prone to overfitting and lack of transparency and interpretation in predictions!

Introduction: GNN Explanation

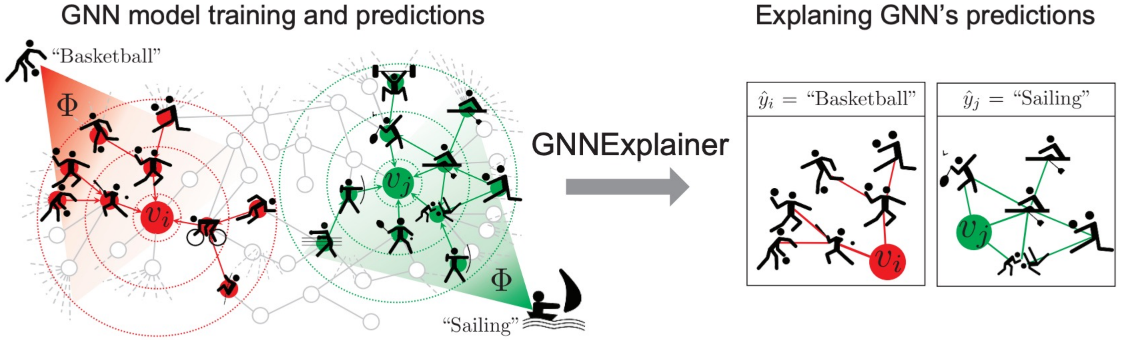
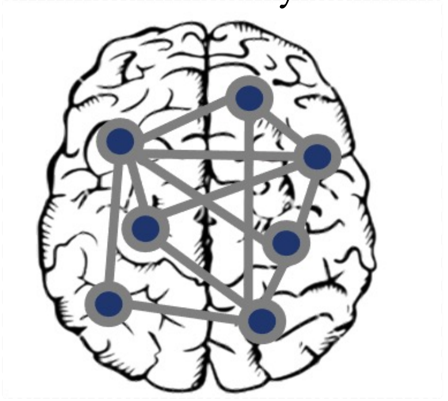


Image source: GNNExplainer: Generating Explanations for Graph Neural Networks

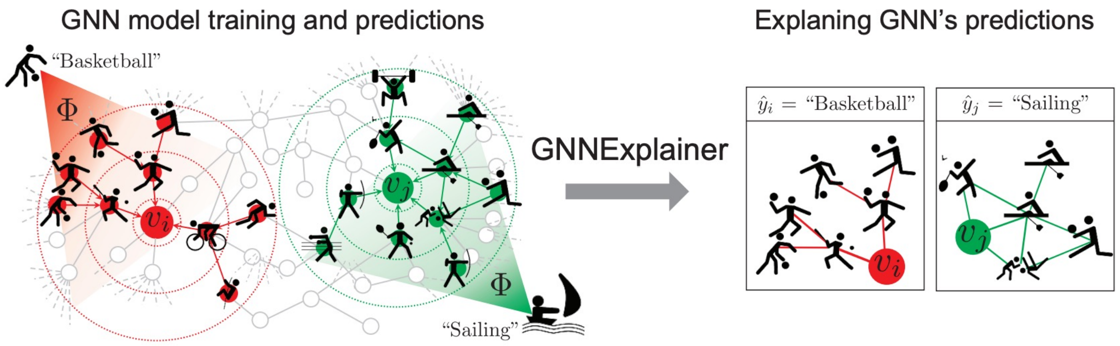
Graph-level connectome-based disorder analysis



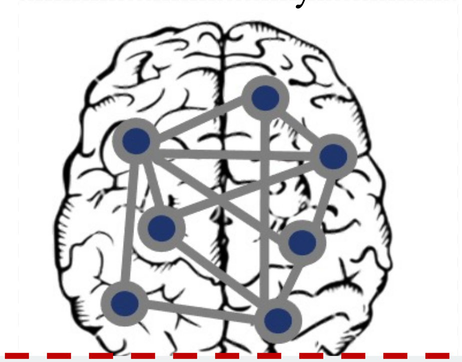
- ❑ Mostly focus on *general graphs* and *node-level* prediction task
- ❑ Produce a *unique explanation for each subject* when applied to graph-level tasks

- ❑ Subjects having the same disorder *share* similar brain network patterns
- ❑ *Unique properties* of brain networks

Introduction: GNN Explanation



Graph-level connectome-based disorder analysis



- ❑ Mostly … *e*
  predict…
- ❑ Produc…
  when a…

Our motivation:

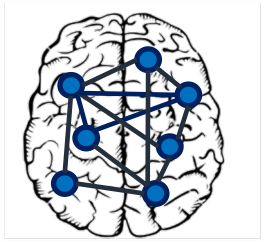
1. **Unleash the *prediction power* of GNNs for brain network analysis**
2. **Investigate *disease-specific patterns* common across the group and provide *interpretations of different levels***

Outline

1. **Introduction & Motivation**
2. **The Proposed Model: IBGNN+** ⬅
    - *The backbone model IBGNN*
    - *The globally shared explanation generator*
    - *Enhancing the backbone with the learned explanations*
3. **Experiments**
4. **Interpretation Analysis**
5. **Conclusion & Future Work**

Problem Definition

- Input: a set of N weighted brain networks, for each network $G = (V, E, \boldsymbol{W})$
  - $V = \{v_i\}_{i=1}^{M}$: Regions Of Interest (ROIs) node set of size M
  - $E = V \times V$: edge set of brain connectome
  - $W \in \mathbb{R}^{M \times M}$: weighted adjacency matrix describing the connection strengths between ROIs



- Output:
  - A brain disorder prediction $\hat{y}_n$ for each subject n
  - A disorder-specific interpretation matrix $\boldsymbol{M} \in \mathbb{R}^{M \times M}$ shared across all subjects, highlighting disorder-specific biomarkers

Overview of the Proposed Model: IBGNN+

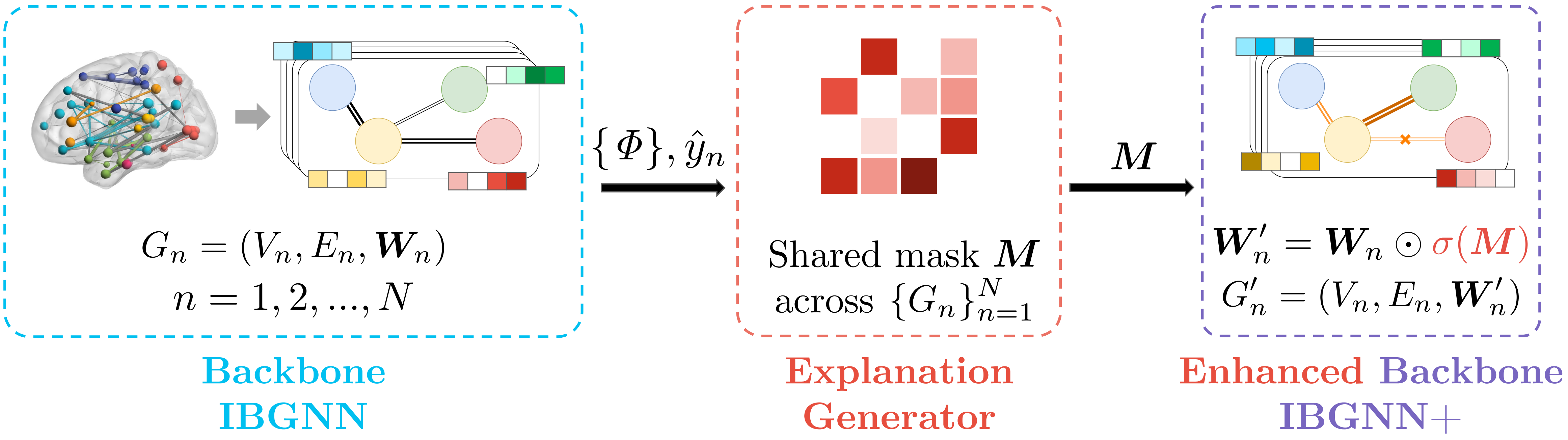


Fig.1: An overview of our proposed framework

- The backbone model is first trained on the original data
- Then, the explanation generator learns a globally shared mask across subjects
- Finally, we enhance the backbone by applying the learned explanation mask and fine-tune the whole model

Module 1: The Backbone Model IBGNN

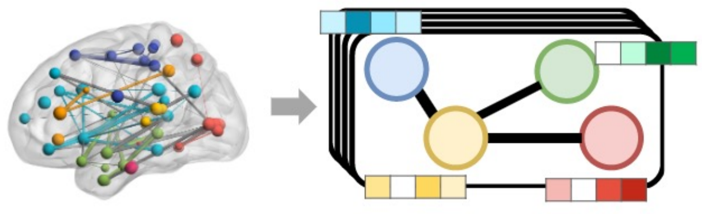
➢ Edge weights in brain networks can be **both positive and negative values**

Message Vector $m_{ij} \in \mathbb{R}^D$: concatenate embeddings of a node v_i and its neighbor v_j, and the edge weight w_{ij}

$$\boldsymbol{m}_{ij}^{(l)} = \mathrm{MLP}_1\left(\left[\boldsymbol{h}_i^{(l)}; \boldsymbol{h}_j^{(l)}; w_{ij}\right]\right)$$

Propagation Rule:

$$\boldsymbol{h}_i^{(l)} = \xi\left(\sum\nolimits_{v_j \in \mathcal{N}_i \cup \{v_i\}} \boldsymbol{m}_{ij}^{(l-1)}\right)$$



$$G_n = (V_n, E_n, \boldsymbol{W}_n)$$
$$n = 1, 2, \ldots, N$$

Backbone IBGNN

Readout Function: summarize all node embeddings to a graph-level one, with MLP and residual connections

$$\boldsymbol{z} = \sum\nolimits_{i \in V} \boldsymbol{h}_i^{(L)} \qquad \boldsymbol{g} = \mathrm{MLP}_2(\boldsymbol{z}) + \boldsymbol{z}$$

***Training Objectives*: supervised cross-entropy**

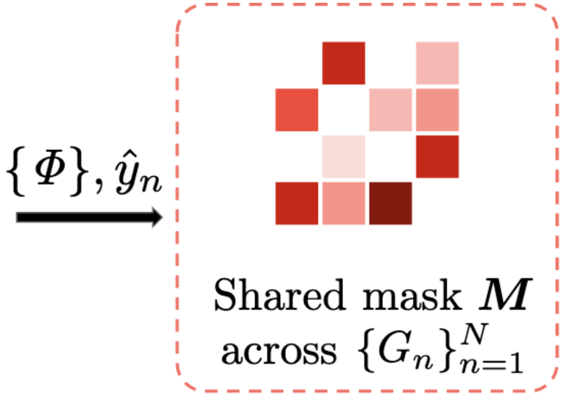
$$\mathcal{L}_{\mathrm{CLF}} = -\frac{1}{N}\sum\nolimits_{n=1}^{N}\left(y_n \log\left(\hat{y}_n\right) + \left(1 - y_n\right)\log\left(1 - \hat{y}_n\right)\right)$$

Module 2: The Globally Shared Explanation Generator

➢ Learn a **globally shared edge mask** $M \in \mathbb{R}^{M \times M}$ that is applied to all brain network subjects in a dataset

Maximize the agreement between the predictions $\hat{y}$ on the original graph G and $\hat{y}'$ on an explanation graph $G' = (V, E, \boldsymbol{W}')$ induced by a masking matrix $\boldsymbol{M}$, where $\boldsymbol{W}' = \boldsymbol{W} \odot \sigma(\boldsymbol{M})$

$$\mathcal{L}_{\text{MASK}} = -\frac{1}{N} \sum_{n=1}^{N} \sum_{c=1}^{C} \mathbb{1}[\hat{y}_n = c] \log P_{\Phi}\left(\hat{y}'_n = \hat{y}_n \mid G'_n\right)$$



Explanation Generator

Two regularization terms: encourage the compactness of the explanation and the discreteness of the mask values

$$\mathcal{L}_{\text{SPS}} = \sum_{i,j} \boldsymbol{M}_{i,j} \qquad \mathcal{L}_{\text{ENT}} = -(\boldsymbol{M} \log(\boldsymbol{M}) + (1 - \boldsymbol{M}) \log(1 - \boldsymbol{M}))$$

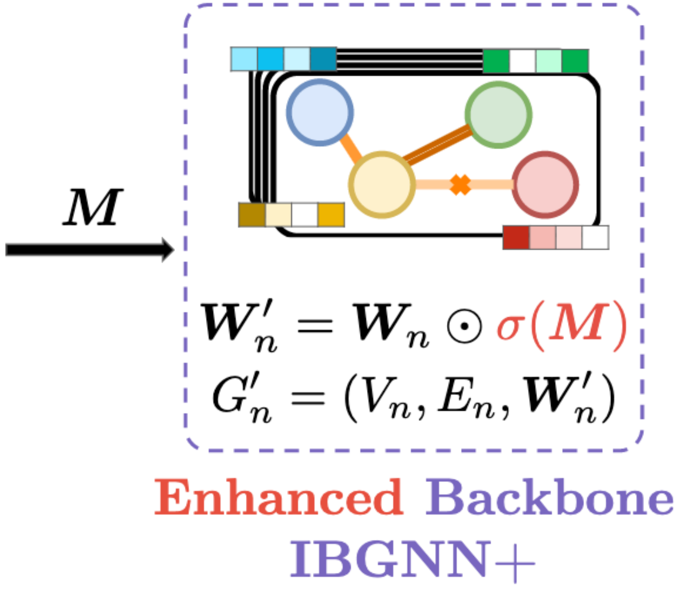
Training Objectives: weighted sum of supervised cross-entropy $\mathcal{L}_{\text{CLF}}$, agreement loss $\mathcal{L}_{\text{MASK}}$, sparsity loss $\mathcal{L}_{\text{SPS}}$, and discreteness loss $\mathcal{L}_{\text{ENT}}$

$$\mathcal{L} = \mathcal{L}_{\text{CLF}} + \alpha \mathcal{L}_{\text{MASK}} + \beta \mathcal{L}_{\text{SPS}} + \gamma \mathcal{L}_{\text{ENT}}$$

Enhancing the Backbone with The Learned Explanation: IBGNN+

- Combine the two modules by applying the shared global explanation mask to individual brain networks →

Predictions and Interpretations are produced in a closed-loop for brain network analysis



Suppress random noises

Highlight essential disorder-specific signals

Outline

1. **Introduction & Motivation**
2. **The Proposed Model: IBGNN+**
   - *The backbone model IBGNN*
   - *The globally shared explanation generator*
   - *Enhancing the backbone with the learned explanations*
3. **Experiments** ⬅
4. **Interpretation Analysis**
5. **Conclusion & Future Work**

Experiments – Datasets

- *Human Immunodeficiency Virus Infection* (HIV)
  preprocessed using DPARSF[1] toolbox

- *Bipolar Disorder* (BP)
  preprocessed using FSL[2] toolbox

- *Parkinson's Progression Markers Initiative* (PPMI)
  preprocessed using FSL[2] toolbox and Advanced Normalization Tools (ANTs)[3]

1 http://rfmri.org/DPARSF/
2 https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/
3 http://stnava.github.io/ANTs/

Dataset	Modality	# samples	Atlas	Size	Response	# classes
HIV	fMRI	70	AAL	90×90	Human Immunodeficiency Virus Infection	2
BP	DTI	97	Brodmann	82×82	Bipolar Disorder	2
PPMI	DTI	754	Desikan-Killiany	84×84	Parkinson's Disorder	2

Tab.1: Dataset summarization

Experiments – Baselines

- Shallow Baselines:
  - M2E[1], MIC[2], MPCA[3], MK-SVM[4]
- Deep Baselines:
  - GCN[5], GAT[6], PNA[7]
- SOTA deep models specifically designed for brain networks:
  - BrainNetCNN[8], BrainGNN[9]

1 Multi-view multi-graph embedding for brain network clustering analysis. AAAI (2018)
2 Clustering on multi-source incomplete data via tensor modeling and factorization. PAKDD (2015)
3 Mpca: Multilinear principal component analysis of tensor objects. IEEE Trans. Neural Netw. (2008)
4 Multimodal analysis of functional and structural disconnection in alzheimer’s disease using multiple kernel svm. Hum. Brain Mapp. (2015)
5 Semi-supervised classification with graph convolutional network. ICLR (2017)
6 Graph attention networks. ICLR (2018)
7 Principal neighbourhood aggregation for graph nets. NeurIPS (2020)
8 Brainnetcnn: Convolutional neural networks for brain networks; towards predicting neurodevelopment. NeuroImage (2017)
9 Braingnn: Interpretable brain graph neural network for fmr ianalysis. Med. Image Anal. (2021)

Experiments – Prediction Performance

Method	HIV			BP			PPMI		
	Accuracy	F1	AUC	Accuracy	F1	AUC	Accuracy	F1	AUC
M2E	57.14±19.17	53.71±19.80	57.50±18.71	52.56±13.86	51.65±13.38	52.42±13.83	78.69±1.78	45.81±4.17	50.39±2.59
MIC	54.29±18.95	53.63±19.44	55.42±19.10	62.67±20.92	63.00±21.61	61.79±21.74	79.11±2.16	49.65±5.10	52.39±2.94
MPCA	67.14±20.25	64.28±23.47	69.17±20.17	52.56±13.12	50.43±14.99	52.42±13.69	79.15±0.57	44.18±0.18	50.00±0.00
MK-SVM	65.71±7.00	62.08±7.49	65.83±7.41	57.00±8.89	41.08±13.44	53.75±8.00	79.15±0.57	44.18±0.18	50.00±0.00
GCN	70.00±12.51	68.35±13.28	73.58±9.49	55.56±13.86	50.71±11.75	61.55±28.77	78.55±1.58	47.87±4.40	59.43±8.64
GAT	71.43±11.66	69.79±10.83	77.17±9.42	63.34±9.15	60.42±7.56	67.07±5.98	79.02±1.25	45.85±3.16	64.40±6.87
PNA	57.14±12.78	45.09±19.62	57.14±12.78	63.71±11.34	55.54±14.06	60.30±11.89	79.36±1.84	51.76±10.32	54.71±6.77
BrainNetCNN	69.24±19.04	67.08±11.11	72.09±19.01	65.83±20.64	64.74±17.42	64.32±13.72	55.20±12.63	55.45±9.15	52.54±10.21
BrainGNN	74.29±12.10	73.49±10.75	75.00±10.56	68.00±12.45	62.33±13.01	74.20±12.93	69.17±0.00	44.19±0.00	45.26±3.65
IBGNN	82.14±10.81*	82.02±10.86*	86.86±11.65*	73.19±12.20	72.87±12.09*	83.64±9.61*	**79.82±1.47**	51.58±4.66	70.65±6.55*
IBGNN+	**84.29±12.94***	**83.86±13.42***	**88.57±10.89***	**76.33±13.00***	**76.13±13.01***	**84.61±9.08***	79.55±1.67	**56.58±7.43**	**72.76±6.73***

Tab.2: Experiment results (%) on three datasets

- Backbone IBGNN outperforms shallow/deep baselines (up to 11% absolute improvement on BP).
- The explanation enhanced IBGNN+ further improve the backbone by ~9.7% relative improvement.
  - ➢ IBGNN+ *can effectively* ***highlight the disorder-specific signals*** *while achieving the benefit of* ***restraining random noises*** *in individual graphs*

Outline

1. **Introduction & Motivation**

2. **The Proposed Model: IBGNN+**
   - *The backbone model IBGNN*
   - *The globally shared explanation generator*
   - *Enhancing the backbone with the learned explanations*

3. **Experiments**

4. **Interpretation Analysis**

5. **Conclusion & Future Work**

Interpretation Analysis – Neural System Mapping

- ROIs on brain networks can be partitioned into different *neural systems*

- 8 commonly used neural systems:
  - Visual Network (VN)
  - Auditory Network (AN)
  - Bilateral Limbic Network (BLN)
  - Default Mode Network (DMN)
  - Somato-Motor Network (SMN)
  - Subcortical Network (SN)
  - Memory Network (MN)
  - Cognitive Control Network (CCN)

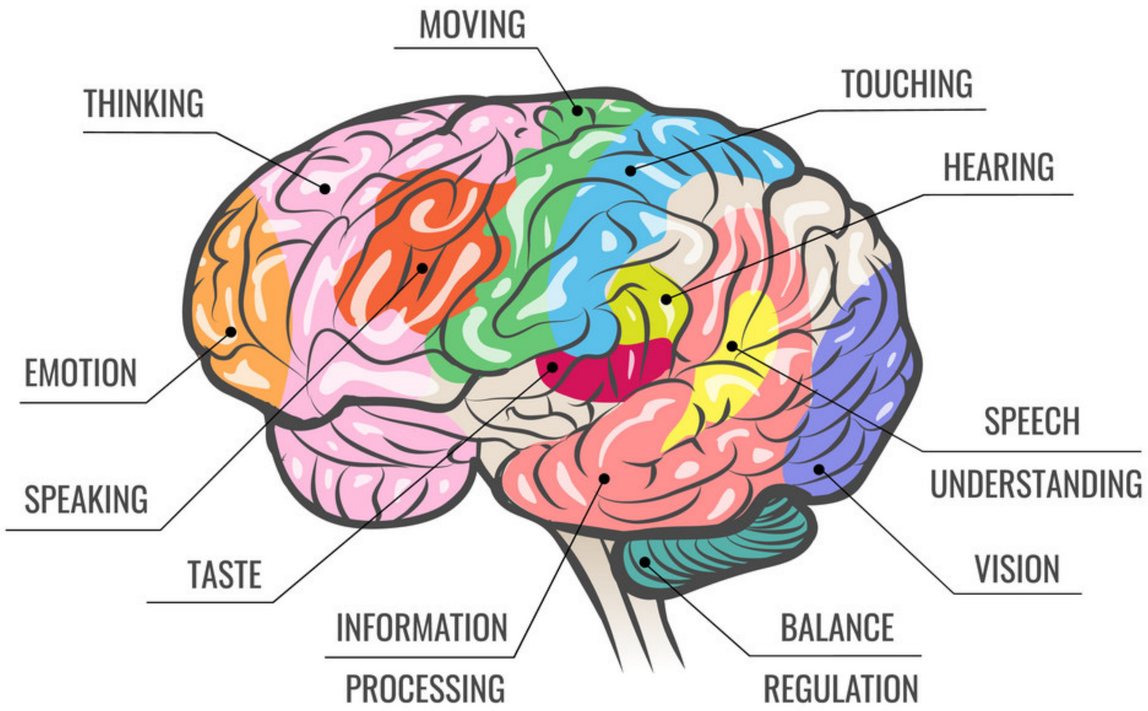


Image source: VectorStock/Human brain function map vector image

Interpretation Analysis – Salient ROIs

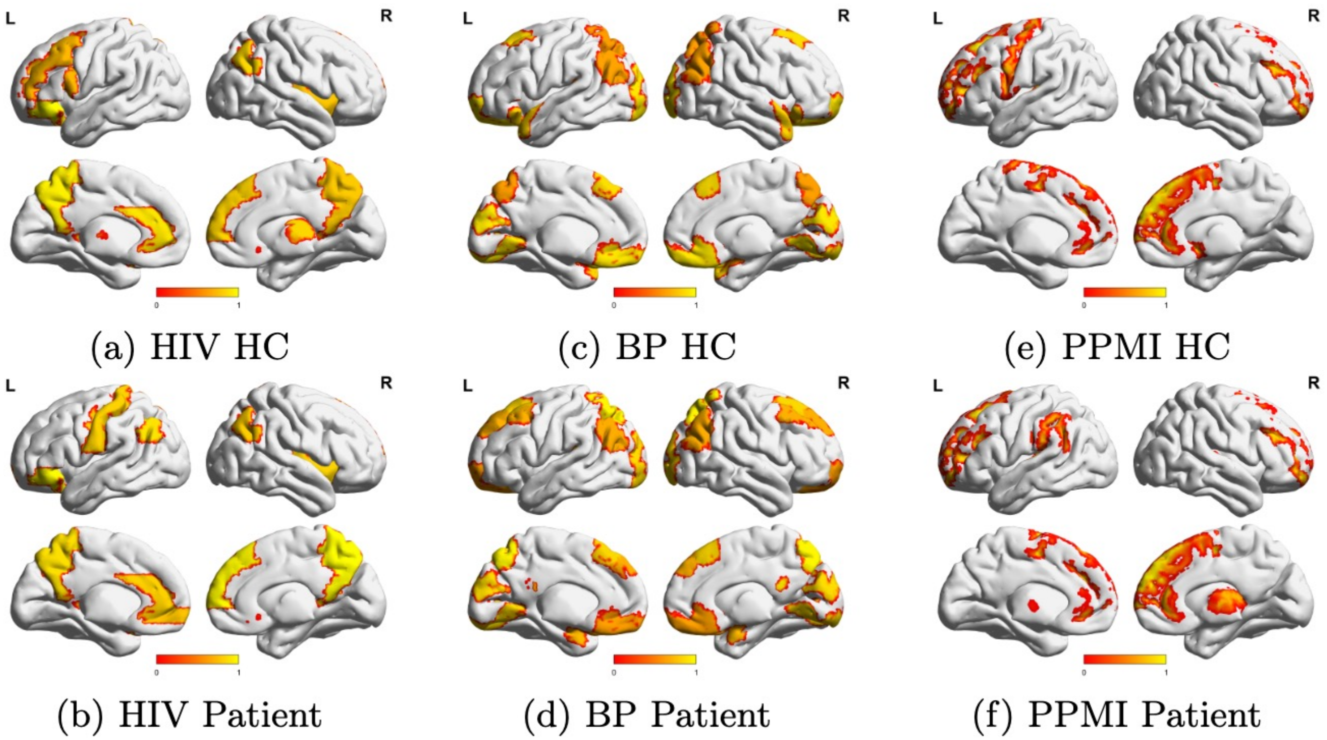


Fig.2: Salient **ROIs** on the explanation enhanced brain connection networks for Health Control (HC) and Patient.

- **Group-level & Individual-level** interpretations on which ROIs contribute most to the prediction of a specific disorder:
  - **HIV**: *anterior cingulate, paracingulate gyri, inferior frontal gyrus*
  - **BP**: *secondary visual cortex and medial to superior temporal gyrus*
  - **PPMI**: *rostral middle frontal gyrus and superior frontal gyrus*
- The observed salient ROIs can be potential **biomarkers** to identify brain disorders from each cohort.

Interpretation Analysis – Important Connections

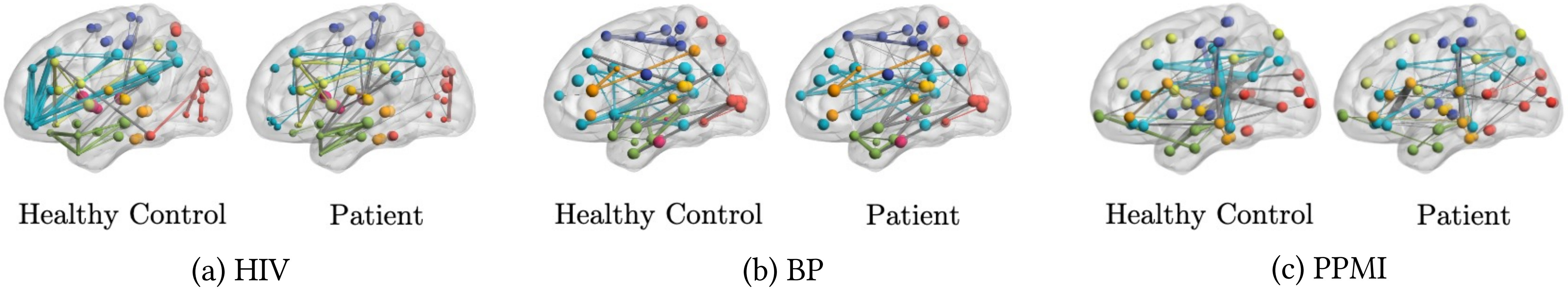


Fig.3: Important **connections** on the explanation enhanced brain connection network. Edges connecting nodes within the same neural system (VN, AN, BLN, DMN, SMN, SN, MN, CCN) are colored accordingly.

Observations:

- **HIV**: patients excludes rich interactions within the DMN and VN systems
- **BP**: connections within BLN system of patients are much sparser
- **PPMI**: connectivity in patients decreases in the SMN and DMN systems

Outline

1. **Introduction & Motivation**
2. **The Proposed Model: IBGNN+**
   - *The backbone model IBGNN*
   - *The globally shared explanation generator*
   - *Enhancing the backbone with the learned explanations*
3. **Experiments**
4. **Interpretation Analysis**
5. **Conclusion & Future Work**

Summary

Conclusion:

- Propose a novel interpretable GNN framework for connectome-based brain disorder analysis
  - *A brain network-oriented GNN predictor*
  - *A globally shared explanation generator*
- Discover disorder-specific interpretations from the generated explanation mask

Future Work:

- Utilize pre-training and transfer learning techniques to learn across datasets and cohorts

Thanks for Listening!

Presenter: Hejie Cui

Paper ***Code***