

# MvHo-IB: Multi-View Higher-Order Information Bottleneck for Brain Disorder Diagnosis

Kunyu Zhang<sup>1</sup>, Qiang Li<sup>2</sup>, and Shujian Yu<sup>3,4</sup>[0000–0002–6385–1705]

<sup>1</sup> International College, Zhengzhou University, 450000 Zhengzhou, China  
kunyu.zky@gmail.com

<sup>2</sup> Tri-Institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS), Georgia State, Georgia Tech, and Emory University, Atlanta, GA 30303, USA

<sup>3</sup> Department of Computer Science, Vrije Universiteit Amsterdam, 1081 HV Amsterdam, The Netherlands

<sup>4</sup> Department of Physics and Technology, UiT - The Arctic University of Norway, 9019 Tromsø, Norway  
yusj9011@gmail.com

**Abstract.** Recent evidence suggests that modeling higher-order interactions (HOIs) in functional magnetic resonance imaging (fMRI) data can enhance the diagnostic accuracy of machine learning systems. However, effectively extracting and leveraging HOIs remains a significant challenge. In this paper, we propose MvHo-IB, a novel multi-view learning framework that seamlessly integrates pairwise interactions and HOIs for diagnostic decision-making while automatically compressing task-irrelevant redundant information. Our approach introduces several key innovations: (1) a principled framework combining  $\mathcal{O}$ -information from information theory with the recently developed matrix-based Rényi’s  $\alpha$ -order entropy functional estimator to quantify and extract HOIs, (2) a purpose-built Brain3DCNN encoder designed to effectively utilize these interactions, and (3) a novel multiview learning information bottleneck objective to enhance representation learning. Experiments on three benchmark fMRI datasets demonstrate that MvHo-IB achieves state-of-the-art performance, outperforming existing methods, including modern hypergraph-based techniques, by significant margins. The code of our MvHo-IB is available at <https://github.com/zky04/MvHo-IB>.

**Keywords:** Multi-view learning · Information Bottleneck ·  $\mathcal{O}$ -information · Matrix-based Rényi’s  $\alpha$ -order entropy functional · fMRI.

## 1 Introduction

Mental disorders exhibit complex neural signatures, making precise neurobiological characterization challenging. Resting-state functional magnetic resonance imaging (rs-fMRI) has emerged as a cornerstone for machine learning-based diagnostic frameworks in mental disorders [14]. With the advent of deep learning, researchers have developed more sophisticated models to analyze brain networks.

Convolutional neural networks primarily captured localized functional connectivity (FC) patterns [8], while graph neural networks (GNNs) further advanced whole-brain analysis by leveraging the complex relational structure of brain networks [15].

However, existing deep learning approaches face two key limitations. First, they predominantly rely on correlations or partial correlations to characterize linear and pairwise FC between brain regions. This fundamentally oversimplifies the role of higher-order interactions (HOIs), which are essential for understanding complex cognitive processes [16]. Second, they overlook the impact of noisy or spurious interactions (i.e., connections influenced by measurement noise or patient-specific artifacts) on the final decision, which may degrade generalization performance and lead to unreliable predictions.

Growing evidence suggests that functional HOIs involving more than two brain regions play a crucial role in neural computation. To model these complex interactions, recent computational approaches have leveraged hypergraph theory [4,26]. Specifically, these methods employ a dual representation scheme, where hyper-nodes map to anatomically segregated brain regions and hyper-edges explicitly encode multivariate functional dependencies [26]. This formulation enables the modeling of concurrent activation patterns across three or more regions, offering a neurobiologically plausible representation of network-level dynamics. However, while hypergraph-based methods offer theoretical advantages for HOI characterization, their practical deployment faces significant bottlenecks. Specifically, hypergraph approaches require manually constructing high-order networks by selecting similarity metrics and pruning rules. The resulting hyperedges, each connecting an arbitrary number of brain regions, merely indicate that these regions are connected, without revealing how they share information (e.g., redundantly or synergistically).

Rather than relying on hypergraphs to model HOIs in a *model-driven* manner, we propose a *data-driven* approach by leveraging  $\mathcal{O}$ -information [18,24] from information theory to capture HOIs. Unlike hyperedges,  $\mathcal{O}$ -information provides a single signed measure that indicates if a set of brain regions generates genuinely new joint information (negative value, synergy-dominated) or primarily reflects repeated signals (positive value, redundancy-dominated). That is,  $\mathcal{O}$ -information not only captures whether regions are connected, as in the case of hyperedges, but also provides a fine-grained quantification of the nature of their interaction. Moreover, constructing HOIs with  $\mathcal{O}$ -information does not require manually selecting similarity metrics or applying pruning rules that may introduce bias.

Our contributions can be summarized as follows:

- We develop MvHo-IB, a novel multi-view learning framework that seamlessly integrates both *nonlinear* pairwise interactions and HOIs while simultaneously eliminating redundant information to enhance predictive performance.
- We propose leveraging  $\mathcal{O}$ -information to capture HOIs and introduce the matrix-based Rényi’s  $\alpha$ -order entropy estimator [32] for its computation. Additionally, we develop Brain3DCNN, a specialized architecture that ex-

- Our MvHo-IB outperforms eight widely used brain network classification methods, demonstrating strong generalization across three datasets while providing clinically interpretable insights that align with clinical evidence.

## 2 Information Bottleneck in Brain Disorder Diagnosis

The IB principle is a framework for extracting the most relevant information from an input variable  $X$  for predicting an output variable  $Y$ . It operates by identifying a “bottleneck” variable  $Z$  that maximizes its predictive power to  $Y$ , as expressed by the mutual information  $I(Y; Z)$ , while imposing some constraints on the amount of information it carries about  $X$ , formulated as  $I(X; Z)$ :

$$\mathcal{L}_{\text{IB}} = I(Y; Z) - \beta I(X; Z), \quad (1)$$

where  $\beta > 0$  is a Lagrange multiplier.

Recent studies have employed the IB principle to enhance both interpretability and generalization in graph-structured data. For example, the dynamic graph attention information bottleneck framework [2] refines raw brain graphs by optimizing graph connectivity and reducing noise, enhancing effective feature aggregation.

The Subgraph Information Bottleneck (SIB) [30] focuses on automatically extracting a predictive subgraph to explain the final decision, thereby enhancing interpretability. Furthermore, BrainIB [34] stabilizes the training of SIB by utilizing the matrix-based Rényi’s  $\alpha$ -order entropy functional, applying this refined framework to the diagnosis of mental disorders.

## 3 Methodology

Consider a dataset of brain signal recordings  $\{X^i, Y^i\}_{i=1}^N$ , where each recording  $X^i \in \mathbb{R}^{C \times T}$  represents the raw blood-oxygen-level-dependent (BOLD) signal, detected in fMRI, for the  $i$ -th patient. Here,  $C$  denotes the number of channels (e.g., 116 for AAL atlas [22] or 105 for ICA-driven brain network template [7]) and  $T$  indicates the signal duration. We further use subscripts to denote the channel index of BOLD signal. Specifically,  $X_i^j$  represents the 1D signal from the  $i$ -th ( $1 \leq i \leq C$ ) brain region for the  $j$ -th ( $1 \leq j \leq N$ ) patient. Our objective is to develop a classifier that maps raw brain signal  $X$  to its label  $Y$ .

### 3.1 Multi-view Information Bottleneck

Figure 1 illustrates the architecture of MvHo-IB. For each participant, we derive two types of brain representations: a  $C \times C$  matrix that encodes all pairwise FCs and a  $C \times C \times C$  3D tensor that captures all three-way interactions, representing the interdependencies among any triplet of brain regions. We treat the raw 2D

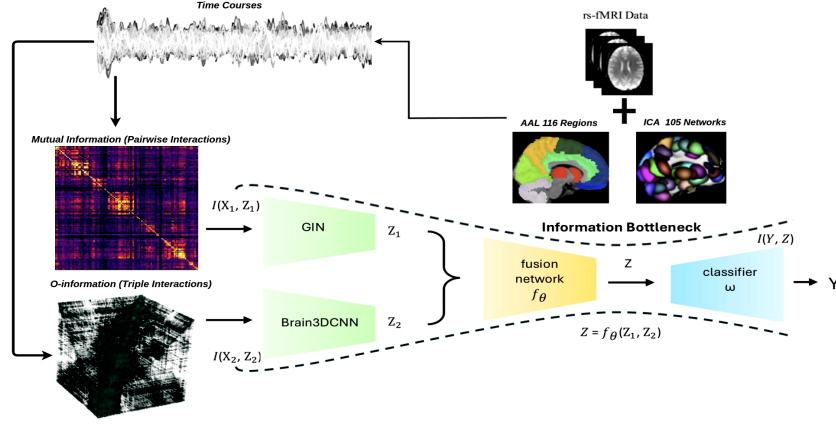


Fig. 1: Illustration of MvHo-IB. The time courses are extracted from fMRI using the automated anatomical labeling (AAL) template [23]. Then, functional connectivity patterns are estimated before being fed into a multi-view framework. The framework learns a joint representation  $Z = f_\theta(Z_1, Z_2)$  by balancing the maximization of  $I(Y; Z)$  with the minimization of  $I(X_1; Z_1) + I(X_2; Z_2)$ , where the first view input is the mutual information matrix (pairwise interactions) and the second view input is the  $\mathcal{O}$ -information 3D tensor (triple interactions).

and 3D representations as complementary views of the raw BOLD signals and propose using a multi-view learning framework to learn a compact and informative joint representation  $z$ . To achieve this, MvHo-IB consists of two encoders  $h_{\phi_1}$  (a Graph Isomorphism Network or GIN to learn  $z_1$ ) and  $h_{\phi_2}$  (a purpose-designed Brain3DCNN that will be detailed in Section 3.3 to learn  $z_2$ ), a feature fusion network  $f_\theta$  to integrate  $z_1$  and  $z_2$  into  $z$ , and a nonlinear classifier  $h_w$ .

Thus, the overall objective of MvHo-IB is formulated as:

$$\begin{aligned} \arg \max_{\phi_1, \phi_2, \theta, w} & \left( I(Y; Z) - \left( \beta_1 I(X_1; Z_1) + \beta_2 I(X_2; Z_2) \right) \right), \\ \text{s.t.} \quad & Z = f_\theta(Z_1, Z_2), \end{aligned} \quad (2)$$

where  $\beta_1$  and  $\beta_2$  are the regularization coefficients for views 1 and 2, respectively.

The prediction term  $I(Z; Y)$  is lower-bounded as follows [10,1]:

$$I(Z; Y) \geq H(Y) + \mathbb{E}_{P(Z, Y)} [\log P(Y | Z)], \quad (3)$$

which essentially optimizes the cross-entropy loss  $\text{CE}(Y, \hat{Y}) = -\mathbb{E}_{P(Z, Y)} [\log P(Y | Z)]$ , since  $H(Y)$  is a constant only depends on input data. On the other hand, for deterministic encoders,  $I(X; Z) = H(Z)$  as the mapping uncertainty  $H(Z|X) =$

0 [19,21]. Consequently, Eq. 2 is reformulated as:

$$\begin{aligned} \arg \min_{\phi_1, \phi_2, \theta, w} & \left( \text{CE}(Y, \hat{Y}) + \beta_1 H(Z_1) + \beta_2 H(Z_2) \right), \\ \text{s.t.} \quad & Z = f_\theta(Z_1, Z_2). \end{aligned} \quad (4)$$

### 3.2 $\mathcal{O}$ -Information and Matrix-based Entropy Functional

We propose utilizing  $\mathcal{O}$ -information [24,18] to capture HOI among any triplet of brain regions. Formally, the  $\mathcal{O}$ -information is defined as the difference between the total correlation (TC) [27] and the dual total correlation (DTC) [6], both of which are nonlinear multivariate dependence measures [33]. The  $\mathcal{O}$ -information for BOLD signals from the  $i$ -th,  $j$ -th, and  $k$ -th brain regions is defined as:

$$\mathcal{O}(X_i, X_j, X_k) = T(X_i, X_j, X_k) - D(X_i, X_j, X_k), \quad (5)$$

if  $\mathcal{O} > 0$ , the triplet is redundancy-dominated, if  $\mathcal{O} < 0$ , synergy dominates [24].

The first term TC  $T(X_i, X_j, X_k)$  is defined as the Kullback-Leibler (KL) Divergence between the joint distribution  $p(X_i, X_j, X_k)$  and product of marginals  $p(X_i)p(X_j)p(X_k)$ , which can be further decomposed as:

$$T(X_i, X_j, X_k) = H(X_i) + H(X_j) + H(X_k) - H(X_i, X_j, X_k), \quad (6)$$

where  $H$  denotes entropy or joint entropy, which can be elegantly estimated with the matrix-based Rényi's  $\alpha$ -order entropy functional [33].

Analogously with TC, the DTC  $D(X_i, X_j, X_k)$  is defined as [6]:

$$D(X_i, X_j, X_k) = H(X_i, X_j, X_k) - H(X_i | X_j, X_k) - H(X_j | X_i, X_k) - H(X_k | X_i, X_j), \quad (7)$$

where  $H(\cdot|\cdot)$  is the conditional entropy.

### 3.3 Brain3DCNN

**3D Edge-to-Edge (E2E) Layer** Inspired by BrainNetCNN [8], our 3D E2E layer processes the 3D tensor by aggregating information from connected edges in a trident kernel along three spatial dimensions (see Figure 2). Formally, let  $\mathbf{O}^{(\ell)} \in \mathbb{R}^{M^\ell \times C \times C \times C}$  denote all feature maps at the  $\ell$ -th layer, extracted from  $M^\ell$  convolutional kernels. For the first input layer,  $\mathbf{O}^1 \in \mathbb{R}^{C \times C \times C}$  is the estimated  $\mathcal{O}$ -information tensor. The output  $\mathbf{O}^{(\ell+1,n)}$  for the  $n$ -th feature map ( $1 \leq n \leq M^{\ell+1}$ ) at the layer  $(\ell + 1)$  is given by:

$$O_{i,j,k}^{(\ell+1,n)} = \sum_{m=1}^{M^\ell} \sum_{c=1}^C \left[ r_d^{(\ell,m,n)} O_{i-c,j,k}^{(\ell,m)} + d_d^{(\ell,m,n)} O_{i,j-c,k}^{(\ell,m)} + e_d^{(\ell,m,n)} O_{i,j,k-c}^{(\ell,m)} \right], \quad (8)$$

where  $r_d^{(\ell,m,n)}$ ,  $d_d^{(\ell,m,n)}$ , and  $e_d^{(\ell,m,n)}$  are learnable weights. Here,  $C$  denotes the spatial extent in each dimension (e.g., 116 for AAL atlas [22] or 105 for ICA-driven brain network template [7]).

**3D Edge-to-Node (E2N) Layer** The 3D E2N layer aggregates the edge-focused representation into a 2D node feature map. Let  $\mathbf{O}^{(\ell,m)} \in \mathbb{R}^{I \times J \times K}$  again be the input feature map; the E2N layer output  $a_i^{(\ell+1,n)}$  for node  $i$  under the  $n$ -th feature map at layer  $(\ell + 1)$  is:

$$a_i^{(\ell+1,n)} = \sum_{m=1}^{M^\ell} \sum_{k=1}^K \left[ r_k^{(\ell,m,n)} O_{i,k}^{(\ell,m)} + d_k^{(\ell,m,n)} O_{k,i}^{(\ell,m)} \right], \quad (9)$$

where  $r_k^{(\ell,m,n)}$  and  $d_k^{(\ell,m,n)}$  are the learnable weights. We apply this 3D-to-2D compression sequentially along spatial dimensions ( $I$ ,  $J$ , or  $K$ ) to reduce the tensor size while preserving key topological relationships between brain regions.

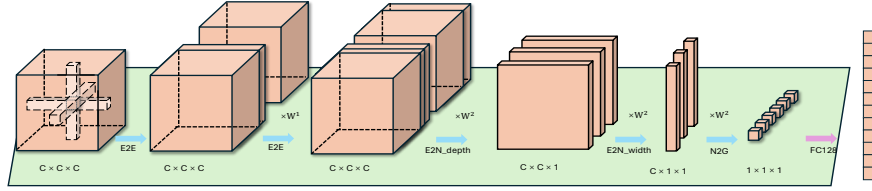


Fig. 2: Each block represents the input/output of filter layers. The brain network adjacency matrix (leftmost block) undergoes E2E convolution, followed by E2N filtering to aggregate edge weights per region. The N2G layer integrates node responses, and fully connected layers refine features for the final prediction.

**3D Node-to-Graph (N2G) Layer** The 3D N2G layer aggregates node features into a scalar. It performs convolution across nodes, capturing spatial relationships in depth, height, and width.

## 4 Experiments and Results

### 4.1 Datasets and Experimental Settings

We evaluate our method on three real-world fMRI datasets. The first dataset, from the UCLA Consortium for Neuropsychiatric Phenomics [5], includes schizophrenia (SZ,  $n = 50$ ) and normal controls (NC,  $n = 114$ ). The second, from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) [11], distinguishes mild cognitive impairment (MCI,  $n = 38$ ) from NC ( $n = 37$ ). Third, the eyes open and eyes closed (EOEC) dataset [35], collected from 48 college students (22 females, aged 19–31), is used for brain-state classification.

All experiments were run on an NVIDIA A100 40G GPU using PyTorch3. The Adam optimizer started with a learning rate of  $1 \times 10^{-5}$ , decaying by 0.5

every 50 epochs, with a weight decay of 0.03. The backbone GIN model uses three GNN layers, each with a two-layer MLP (hidden dimensions: [128, 256, 512]), followed by batch normalization and ReLU activation. The Brain3DCNN includes E2E3D layers (32 to 64 channels) and spatial-channel reduction modules. The model was trained for 100 epochs with a batch size of 32 and a dropout rate of 0.5. We set the matrix-based Rényi’s  $\alpha$ -order entropy hyperparameters to  $\sigma = 5$  and  $\alpha = 1.01$  [33,32]. Regularization coefficients  $\beta_1$  and  $\beta_2$  were tuned over  $\{0, 0.0001, 0.001, 0.01, 0.1\}$  using tenfold cross-validation, with MvHo-IB achieving the best accuracy at  $\beta_1 = 0.01$  and  $\beta_2 = 0.1$ . The fusion module combined both views via a 3-layer MLP with ReLU and dropout ( $p = 0.5$ ). All competing models were trained with their recommended hyperparameters.

## 4.2 Experimental Results and Ablation Study

We compare our proposed model with eight different methods, including three representative GNN models: GCN [9], GAT [25] and GIN [29], three state-of-the-art approaches based on information-theoretic principles: SIB [31], DIR-GNN [28] and BrainIB [34], and two hypergraph-based approaches: HYBRID [17] and HMNet [13], in three datasets in Table 1. As can be seen, our approach consistently outperforms others by a significant margin, particularly on the UCLA and ADNI. BrainIB and hypergraph-based methods follow in ranking, leveraging information compression and HOIs, respectively, to enhance generalization. By naturally integrating both merits, our approach achieves the best performance.

Table 1: Tenfold cross-validation performances of different models. The best performance is in bold, and the second-best is underlined.

Method	UCLA	ADNI	EOEC
GCN [9]	62.27 $\pm$ 6.21	66.13 $\pm$ 4.62	70.92 $\pm$ 8.56
GAT [25]	67.73 $\pm$ 7.61	66.28 $\pm$ 8.69	72.73 $\pm$ 8.64
GIN [29]	65.91 $\pm$ 8.21	68.33 $\pm$ 6.47	75.41 $\pm$ 9.65
DIR-GNN [28]	75.72 $\pm$ 8.37	70.63 $\pm$ 6.96	80.12 $\pm$ 6.21
SIB [31]	72.76 $\pm$ 8.13	70.12 $\pm$ 7.43	80.42 $\pm$ 7.97
BrainIB [34]	79.14 $\pm$ 4.17	72.47 $\pm$ 5.32	82.06 $\pm$ 5.43
HYBRID [17]	<u>79.38 <math>\pm</math> 8.34</u>	<u>71.34 <math>\pm</math> 7.43</u>	81.97 $\pm$ 7.43
MHNet [13]	79.22 $\pm$ 6.72	71.96 $\pm$ 4.96	<u>82.87 <math>\pm</math> 5.43</u>
MvHo-IB	<b>83.12 <math>\pm</math> 5.74</b>	<b>73.23 <math>\pm</math> 4.37</b>	<b>82.13 <math>\pm</math> 6.96</b>

To clarify the role of each module in our framework, we conduct an ablation study: Use only the GIN with pairwise FC measured with Pearson’s correlation coefficient  $\rho$ . Combine GIN with nonlinear pairwise interaction measured with mutual information  $I_\alpha$ , while excluding the  $\mathcal{O}$ -information 3D tensor. Integrate the  $\mathcal{O}$ -information 3D tensor into the Brain3DCNN pathway. Enable both GIN

and Brain3DCNN while incorporating the mutual information-based FC and the  $\mathcal{O}$ -information 3D tensor. Finally, include all components of MvHo-IB, together with the information bottleneck regularization, i.e., the last two terms in Eq. 4.

Table 2: Tenfold cross-validation results for the ablation study on three datasets. Bold indicates the best performance, while underlined denotes the second-best.

Dataset	$\rho$	$I_\alpha$	$\mathcal{O}_\alpha$	$I_\alpha + \mathcal{O}_\alpha$ (w/o IB)	$I_\alpha + \mathcal{O}_\alpha$ (w. IB)
UCLA	$65.91 \pm 8.21$	$73.26 \pm 8.43$	$74.29 \pm 5.43$	<u><math>82.51 \pm 5.96</math></u>	<b><math>83.12 \pm 5.74</math></b>
ADNI	$68.33 \pm 6.47$	$69.23 \pm 7.29$	$70.81 \pm 6.02$	<u><math>72.21 \pm 6.10</math></u>	<b><math>73.23 \pm 4.37</math></b>
EOEC	$75.41 \pm 9.65$	$76.52 \pm 6.61$	$77.63 \pm 6.17$	<u><math>81.34 \pm 6.92</math></u>	<b><math>82.13 \pm 6.96</math></b>

As shown in Table 2, both the mutual information nonlinear pairwise interaction and the three-way  $\mathcal{O}$ -information HOI contain more discriminative information than simply using  $\rho$ . The multiview learning framework with IB regularization naturally fuses two kinds of complementary information, while effectively removing redundant information, thereby improving generalization.

Table 3: The top two interpretable pairwise and three-way interactions used by our model, identified with Grad-CAM. For EOEC, the top two pairwise groups are identical, only one group is presented.

Dataset	pairwise	three-way HOI
UCLA	HC-HC	HC-HC-TP
	HC-SC	HC-HC-SM
ADNI	SMN-FPN	SMN-FPN-FPN
	FPN-DMN	SMN-FPN-CON
EOEC	SMN-FPN	SMN-FPN-FPN
		FPN-DMN-CON

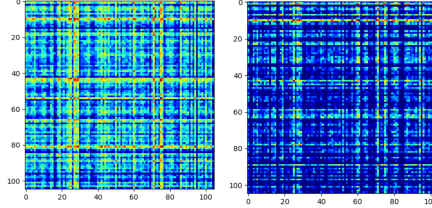


Fig. 3: Heatmap visualization of interpretable connectivities by Grad-CAM in UCLA dataset. Left: pairwise interactions; Right: one slice of 3D tensor interactions.

In order to validate our model’s interpretability, we applied Grad-CAM [20] to three datasets as shown in Figure 3. The abbreviations are: sensorimotor network (SMN), occipital network (ON), fronto-parietal network (FPN), default mode network (DMN), cingulo-opercular network (CON), cerebellum network (CN), Higher Cognition network (HC), Subcortical network (SC), Temporal network (TP), and Sensorimotor network (SM). We generated pairwise and three-way Grad-CAM heatmaps in two input views to identify the discriminative interactions distinguishing healthy individuals from those with mental illness. Table 3 shows informative pairwise interactions within HC and between HC and



SC in the UCLA schizophrenia dataset, aligning with findings from previous studies [12]. Moreover, our model revealed significant and novel triple-network interactions among HC, TP, and SM, beyond pairwise associations, underscoring the added value of triple-network analysis for future research and validation.

## 5 Conclusions and Future Work

In this work, we propose a novel and effective framework that leverages HOIs to enhance diagnostic accuracy in fMRI-based mental disorder classification. Our method serves as a principled alternative to hypergraphs. We validate our framework on three benchmark fMRI datasets and compare it against eight competitive methods, confirming its superior precision. Moreover, interpretability analyses verify the reliability of our approach by revealing neurobiologically plausible three-way HOI biomarkers that offer new and promising insights into the distributed neural mechanisms underlying psychiatric conditions.

While this paper focuses on third-order  $\mathcal{O}$ -information, the formulation naturally extends to higher orders, yielding a  $K$ -way tensor [12]. However, computing higher-order  $\mathcal{O}$ -information is computationally demanding. Two promising directions for scalability are: (1) leveraging an analytical form under Gaussian assumptions [12], and (2) using low-rank approximations of the matrix-based entropy functional [3].

**Acknowledgments.** This study was funded in part by the Research Council of Norway (RCN) under grant 309439.

**Disclosure of Interests.** No competing interests.

## References

1. Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017.
2. Changxu Dong and Dengdi Sun. Brain network classification based on dynamic graph attention information bottleneck. *Computer Methods and Programs in Biomedicine*, 243:107913, 2024.
3. Yuxin Dong, Tieliang Gong, Shujian Yu, and Chen Li. Optimal randomized approximations for matrix-based rényi’s entropy. *IEEE Transactions on Information Theory*, 69(7):4218–4234, 2023.
4. Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3558–3565, Jul. 2019.
5. Krzysztof Gorgolewski, Joke Durnez, and Russell Poldrack. Preprocessed consortium for neuropsychiatric phenomics dataset. *F1000Research*, 6:1262, 09 2017.
6. Te Sun Han. Nonnegative entropy measures of multivariate symmetric correlations. *Inf. Control.*, 36:133–156, 1978.

7. Armin Iraj et al. Identifying canonical and replicable multi-scale intrinsic connectivity networks in 100k+ resting-state fmri datasets. *Human Brain Mapping*, 44, 10 2023.
8. J. Kawahara, C. J. Brown, S. P. Miller, et al. Brainnetcnn: Convolutional neural networks for brain networks; towards predicting neurodevelopment. *NeuroImage*, 146:1038–1049, 2017.
9. Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.
10. Artemy Kolchinsky, Brendan D. Tracey, and David H. Wolpert. Nonlinear information bottleneck. *Entropy*, 21(12):1181, November 2019.
11. Liqun Kuang, Xie Han, Kewei Chen, Richard J. Caselli, Eric M. Reiman, and Yalin Wang. A concise and persistent feature to study brain resting-state network dynamics: Findings from the alzheimer’s disease neuroimaging initiative. *Human Brain Mapping*, 40:1062 – 1081, 2018.
12. Qiang Li, Vince Calhoun, Adithya Ram Ballem, Shujian Yu, Jesus Malo, and Armin Iraj. Aberrant high-order dependencies in schizophrenia resting-state functional MRI networks. In *NeurIPS 2023 workshop: Information-Theoretic Principles in Cognitive Systems*, 2023.
13. Yueyang Li et al. Mhnet: Multi-view high-order network for diagnosing neurodevelopmental disorders using resting-state fmri. *Journal of imaging informatics in medicine*, 2024.
14. M. Marshall. The hidden links between mental disorders. *Nature*, 581(7806):19–22, 2020.
15. G. Märtensson, J. B. Pereira, P. Mecocci, B. Vellas, M. Tsolaki, I. Ktoszewska, H. Soininen, S. Lovestone, A. Simmons, G. Volpe, et al. Stability of graph theoretical measures in structural brain networks in alzheimer’s disease. *Scientific Reports*, 8(1):1–15, 2018.
16. P. Prado, A. Ibanez, et al. Genuine high-order interactions in brain networks and neurodegeneration. *Neurobiology of Disease*, 165:105918, 2022. Available online 12 November 2022.
17. Weikang Qiu, Huangrui Chu, Selena Wang, Haolan Zuo, Xiaoxiao Li, Yize Zhao, and Rex Ying. Learning high-order relationships of brain regions. *ArXiv*, abs/2312.02203, 2023.
18. Fernando E. Rosas, Pedro A. M. Mediano, Michael Gastpar, and Henrik Jeldtoft Jensen. Quantifying high-order interdependencies via multivariate extensions of the mutual information. *Physical review. E*, 100 3-1:032305, 2019.
19. Andrew Michael Saxe et al. On the information bottleneck theory of deep learning. In *International Conference on Learning Representations*, 2018.
20. Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128:336 – 359, 2016.
21. Daniel J Strouse and David J Schwab. The deterministic information bottleneck. *Neural computation*, 29(6):1611–1630, 2017.
22. Nathalie Tzourio-Mazoyer et al. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *NeuroImage*, 15:273–89, 02 2002.
23. Nathalie Tzourio-Mazoyer, Brigitte Landeau, Dimitri Papathanassiou, Fabrice Crivello, Octave Etard, Nicolas Delcroix, Bernard Mazoyer, and Marc Joliot. Au-

- tomated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *Neuroimage*, 15(1):273–289, 2002.
24. Thomas F. Varley, Maria Pope, Joshua Faskowitz, and Olaf Sporns. Multivariate information theory uncovers synergistic subsystems of the human cerebral cortex. *Communications Biology*, 6(1):Article 4843, 2023.
  25. Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.
  26. Wei Wang, Li Xiao, Gang Qu, Vince D. Calhoun, Yu-Ping Wang, and Xiaoyan Sun. Multiview hyperedge-aware hypergraph embedding learning for multisite, multiatlas fmri based functional connectivity network analysis. *Medical Image Analysis*, 94:103144, 2024.
  27. Michael Satoshi Watanabe. Information theoretical analysis of multivariate correlation. *IBM J. Res. Dev.*, 4:66–82, 1960.
  28. Yingxin Wu, Xiang Wang, An Zhang, Xiangnan He, and Tat-Seng Chua. Discovering invariant rationales for graph neural networks. In *International Conference on Learning Representations*, 2022.
  29. Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *ArXiv*, abs/1810.00826, 2018.
  30. J. Yu, T. Xu, Y. Rong, Y. Bian, J. Huang, and R. He. Recognizing predictive substructures with subgraph information bottleneck. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.
  31. Junchi Yu, Tingyang Xu, Yu Rong, Yatao Bian, Junzhou Huang, and Ran He. Recognizing predictive substructures with subgraph information bottleneck. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(3):1650–1663, 2024.
  32. Shujian Yu et al. Multivariate extension of matrix-based rényi’s  $\alpha$ -order entropy functional. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
  33. Shujian Yu et al. Measuring dependence with matrix-based entropy functional. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10781–10789, 2021.
  34. Kaizhong Zheng et al. Brainib: Interpretable brain network-based psychiatric diagnosis with graph information bottleneck. *IEEE Transactions on Neural Networks and Learning Systems*, page 1–14, 2024.
  35. Zhen Zhou et al. A toolbox for brain network construction and classification (brainnetclass). *Human Brain Mapping*, 41(4):3832–3848, 2020.