Learning High-Order Relationships of Brain Regions

Weikang Qiu [1] Huangrui Chu [1] Selena Wang [1] Haolan Zuo [1] Xiaoxiao Li [2 3] Yize Zhao [1] Rex Ying [1]

Abstract

Discovering reliable and informative relationships among brain regions from functional magnetic resonance imaging (fMRI) signals is essential in phenotypic predictions. Most of the current methods fail to accurately characterize those interactions because they only focus on pairwise connections and overlook the high-order relationships of brain regions. We propose that these high-order relationships should be *maximally informative and minimally redundant* (MIMR). However, identifying such high-order relationships is challenging and under-explored due to the exponential search space and the absence of a tractable objective. In response to this gap, we propose a novel method named HyBRiD which aims to extract MIMR high-order relationships from fMRI data. HyBRiD employs a CONSTRUCTOR to identify hyperedge structures, and a WEIGHTER to compute a weight for each hyperedge, which avoids searching in exponential space. HyBRiD achieves the MIMR objective through an innovative information bottleneck framework named *multi-head drop-bottleneck* with theoretical guarantees. Our comprehensive experiments demonstrate the effectiveness of our model. Our model outperforms the state-of-the-art predictive model by an average of 11.2%, regarding the quality of hyperedges measured by CPM, a standard protocol for studying brain connections. Source code is available at https://github.com/Graph-and-Geometric-Learning/HyBRiD.

1. Introduction

Discovering relations among brain regions toward a specific phenotypic outcome from fMRI signals has been a crucial area in neuroimaging research. Reliable and informative relations help neuroscientists and clinical professionals to better understand brain functions, and thus improve clinical diagnosis and treatments (Kucian et al., 2008; 2006; Li et al., 2015b;a; Satterthwaite et al., 2015; Wang et al., 2016). However, despite the clear multiplexity of the brain's involvement in cognition (Logue & Gould, 2014; Barrasso-Catanzaro & Eslinger, 2016; Knauff & Wolf, 2010; Reineberg et al., 2022), current imaging biomarker detection methods (Shen et al., 2017; Gao et al., 2019; Li et al., 2021) focus only on the contributing roles of the pairwise connectivity edges. In contrast, most brain functions involve distributed patterns of interactions among multiple regions (Semedo et al., 2019). For instance, executive planning requires the appropriate communication of signals across many distinct cortical areas (Logue & Gould, 2014). These high-order relationships, cannot always be decomposed into pairwise ones (Battiston et al., 2020; 2021; Bick et al., 2023), thus allowing them to capture information beyond the reach of pairwise ones (Do et al., 2020). Consequently, relying solely on pairwise connectivity, without accounting for the brain's complex high-order structure, may result in inconsistent findings and low predictive performance across studies. Although recently there have been a few works (Zu et al., 2016; Xiao et al., 2019; Li et al., 2022) working on discovering the high-order relationships of brain regions, they are unable to effectively extract meaningful patterns. This is because these methods usually first identify some candidates of high-order relationships and then perform feature selection. The candidates are obtained by enumerating only a small portion of all possible relations, or by clustering methods that are unrelated to the target. As a result, most informative high-order relations are likely excluded at the first stage. This inspires us to solve the problem in an end-to-end manner through a more expressive model and a more appropriate objective.

In this paper, we aim to identify high-order relationships that are informative towards a phenotypic outcome, such as a cognition score. However, unlike pairwise relations, the number of possible high-order relations is exponential. To identify the most informative ones from the exponential

[1]Yale University, New Haven, USA [2]University of British Columbia, Vancouver, Canada [3]Vector Institute, Toronto, Canada. Correspondence to: Rex Ying <rex.ying@yale.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

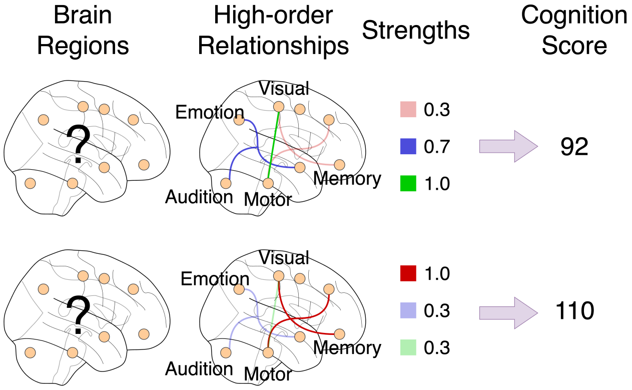


Figure 1: We identify high-order relationships of brain regions, where hyperedge structures and weights possess strong relevance to cognition (maximal informativeness). Meanwhile, they contain the least irrelevant information.

space, we propose our objective: *maximally informative and minimally redundant* (MIMR). That is to say, we maximize the information contained in the high-order relationships towards a neurological outcome (*informativeness*) while diminishing the participation of unrelated brain regions (*redundant*). Such a criterion, on the one hand, ensures the predictive performance of these high-order relationships; on the other hand, it endows the model with the capacity to identify more succinct and interpretable structures (Yu et al., 2020; Miao et al., 2022a;b; Chen & Ying, 2023). A formal definition of the MIMR criterion could be found in Equation 8 from an information bottleneck point of view.

We formulate high-order relationships as weighted hyperedges in a hypergraph, where regions are treated as nodes. Unlike a traditional graph where edges connect only two nodes, a hypergraph allows edges, known as hyperedges, to connect any number of nodes. The hypergraph should be weighted, and the weights of hyperedges are considered as strengths of high-order relationships, which contain the information relevant to the outcome (Figure 1).

However, current methods for hypergraph construction, which are mostly based on neighbor distances and neighbor reconstruction (Wang et al., 2015; Liu et al., 2017; Jin et al., 2019; Huang et al., 2009), are unsuitable in our context for several reasons: 1) they are unable to learn MIMR hyperedges due to the absence of a tractable objective for learning such hyperedges. 2) they fall short of learning consistent structures across subjects, which contradicts the belief that the cognitive function mechanism of healthy human beings should be similar (Wang et al., 2023b). 3) the number of hyperedges is restricted to the number of nodes, which may lead to sub-optimal performance. Furthermore, although information bottleneck (IB) has been a prevailing solution to learn MIMR representations in deep learning (Kim et al., 2021; Alemi et al., 2016; Luo et al., 2019), existing IB methods focus on extracting compressed representations of inputs instead of identifying underlying structures such as hypergraphs. Harnessing the IB framework for identifying hypergraphs necessitates both architectural innovations and theoretical derivations.

Proposed Work In this paper, we propose **Hy**pergraph of **B**rain **R**egions via mult**i**-head **D**rop-bottleneck (HyBRId), a novel approach for identifying maximally informative yet minimally redundant high-order relationships of brain regions. The overall pipeline of HyBRId is depicted in Figure 2. HyBRId is equipped with a Constructor and a Weighter. The Constructor identifies the hyperedge structures of brain regions by learning sets of masks, and the Weighter computes a weight for each hyperedge. To advance the IB principle for hyperedge identification, we further propose *multi-head drop-bottleneck* and derive its optimization objective.

HyBRId avoids searching in an exponential space through learning masks to identify hyperedges, which guarantees efficiency. Its feature-agnostic masking mechanism design ensures HyBRId to learn consistent structures across subjects. Moreover, the model is equipped with a number of parallel heads, each of which is dedicated to a hyperedge. Through this, HyBRId is able to identify any number of hyperedges, depending on how many heads it is equipped with. Additionally, the proposed *multi-head drop-bottleneck* theoretically guarantees the maximal informativeness and minimal redundancy of the identified hyperedges.

We evaluate our methods on the open-source ABIDE dataset and the restricted ABCD dataset. We quantitatively evaluate our approach by a commonly used protocol for studying brain connections, CPM (Shen et al., 2017) (Appendix B), and show that our model outperforms the state-of-the-art deep learning models by an average of 11.2% on a comprehensive benchmark. Our post-hoc analysis demonstrates that hyperedges of higher degrees are considered more significant, which indicates the significance of high-order relationships in human brains.

2. Problem Definition & Notations

Input Our dataset is a collection of human subject's features and their phenotypic outcomes, which is represented by the pair (X, Y) for each subject. $X \in \mathbb{R}^{N \times d}$ represents the features for each subject, where N is the number of brain regions and d is the feature size. Consistent with previous works (Kan et al., 2022b; Li et al., 2021), the features are Pearson correlations derived from fMRI time series. $Y \in \mathbb{R}$ denotes the phenotypic outcome, such as the intelligent quotient. Section 5 and Appendix D will elaborate more details about datasets and preprocessing procedures.

Goal Based on the input X, HyBRiD aims to learn a weighted hypergraph of the brain, where regions are nodes. To achieve this, HyBRiD identifies a collection of hyperedges $H = (\boldsymbol{h}^1, \boldsymbol{h}^2, \cdots, \boldsymbol{h}^K)$, and assigns weights $\boldsymbol{w} = [w^1, w^2, \cdots, w^K]^T$ for all hyperedges. These hyperedges and their weights, which represent strengths of hyperedges, are expected to be the most informative towards Y yet the least redundant.

Representation of Hyperedges As mentioned before, we use H to denote the collection of hyperedge structures and $\boldsymbol{h}^k$ to denote the k-th hyperedge. To associate hyperedges with node memberships, we use the following representation for a hyperedge:

$$\boldsymbol{h}^k = \boldsymbol{m}^k \odot X \in \mathbb{R}^{N\times d}, \tag{1}$$

where $\boldsymbol{m}^k \in \{0,1\}^N$ is a mask vector and $\odot$ denotes broadcasting element-wise multiplication. In other words, each $\boldsymbol{h}^k$ is a randomly row-zeroed version of X.

3. Related Work

Hypergraph Construction Existing hypergraph construction methods are mostly based on neighbor reconstruction and neighbor distances. For example, the k nearest neighbor-based method (Huang et al., 2009) connects a centroid node and its k nearest neighbors in the feature space to form a hyperedge. Wang et al. (2015); Liu et al. (2017); Jin et al. (2019); Xiao et al. (2019) further refine these neighbor connections through various regularization. However, the number of hyperedges of these methods is restricted to the number of nodes, and hyperedges obtained by these methods are inconsistent across instances. Cluster-based approaches learn community structure in a graph (Bannadabhavi et al., 2023; Ying et al., 2018), which can also be used to form hyperedges. Zhang et al. (2022; 2018) proposed to iteratively refine a noisy hypergraph, which is obtained by the aforementioned methods. Therefore, they share the same limitations as the aforementioned methods. In addition, these methods are unable to learn MIMR hyperedges due to the absence of a tractable objective. Other methods, such as attributed-based methods (Huang et al., 2015; Joslyn et al., 2019), are ill-adapted to our context since they require discrete labels. Different from these methods, we provide a way to learn a consistent hypergraph through a deep-learning model. Furthermore, thanks to the proposed *multi-head drop-bottleneck*, these hyperedges are theoretically ensured MIMR.

High-Order Relationships in fMRI Although there are some methods working on high-order relationships in fMRI, they have limitations and are inconsistent with our MIMR objective. Xiao et al. (2019); Li et al. (2022) used the existing non-learning-based hypergraph construction methods, which may lead to noisy and inexpressive hypergraphs. Zu et al. (2016); Santoro et al. (2023) enumerated all hyperedges with degrees lower than 3, which can only discover a tiny portion of all possible hyperedges in exponential space and is not scalable to a large degree. Rosas et al. (2019) proposed O-information, which reflects the balance between redundancy and synergy. The O-information metric is utilized by Varley et al. (2023) to study fMRI data. However, the objective of these methods is not consistent with ours: although both of us are quantifying the redundancy of high-order relations, our method is to learn those that are most informative toward a cognition score, while theirs is to depict the synergy and redundancy within a system.

Information Bottleneck Information bottleneck (IB) (Tishby et al., 2000) is a technique in data compression. The key idea is to extract a summary of data, which contains the most relevant information to the objective. Alemi et al. (2016) first employed an IB view of deep learning. After that, IB has been widely used in deep learning. The applications span areas such as computer vision (Luo et al., 2019; Peng et al., 2018), reinforcement learning (Goyal et al., 2019; Igl et al., 2019), natural language processing (Wang et al., 2020) and graph learning (Yu et al., 2020; 2022; Xu et al., 2021; Wu et al., 2020). Unlike these studies that use IB to extract a compressed representation or a select set of features, our approach focuses on identifying the underlying structures of the data.

Connectivity-based Phenotypic Prediction Recently, deep learning techniques have been increasingly employed in predicting phenotypic outcomes based on the connectivity of brain regions. Most works (Ahmedt-Aristizabal et al., 2021; Li et al., 2019; Cui et al., 2022b; Kan et al., 2022a; Cui et al., 2022a; Said et al., 2023) model the brain network as a graph, in which regions act as nodes and pairwise correlations form the edges. These methods predominantly utilize Graph Neural Networks (GNNs) to capture the connectivity information for predictions. In addition to GNNs, Kan et al. (2022b) proposed to use transformers with a specially designed readout module, leveraging multi-head attention mechanisms to capture pairwise connectivity. However, all of these methods heavily rely on pairwise connectivity and neglect more intricate higher-order relationships. This oversight, on the one hand, leads to sub-optimal prediction performances and, on the other hand, prevents domain experts from acquiring insightful neuroscience interpretations, given that brain functions often involves multiple regions.

4. Method

Method Overview HyBRiD consists of a Constructor $\mathcal{F}_c$, a Weighter $\mathcal{F}_w$, and a LinearHead $\mathcal{F}_l$. At a high level, the Constructor $\mathcal{F}_c$ is responsible for identifying hyperedges H from the data to construct the hypergraph. After that, the WEIGHTER $\mathcal{F}_w$ calculates a weight for each hyperedge. Finally, based on all the weights $\boldsymbol{w}$, the LINEARHEAD $\mathcal{F}_l$ predicts the label Y. An illustration of this pipeline is shown in Figure 2. The pipeline can be formulated as

$$X \underset{\mathcal{F}_c}{\longrightarrow} H \underset{\mathcal{F}_w}{\longrightarrow} \boldsymbol{w} \underset{\mathcal{F}_l}{\longrightarrow} Y. \tag{2}$$

We will elaborate on the details of the architecture below.

4.1. Learning the Hypergraph by Multi-head Masking

Each instance (human subject) is represented by $X = [X_1, X_2, \cdots, X_N]^T \in \mathbb{R}^{N \times d}$, where $X_i \in \mathbb{R}$ is a column vector representing the features of region i. These regions are nodes in the hypergraph we are going to construct. Hyperedges in the hypergraph can be beneficial in the learning process below because it is essential to model the relationships between more than two regions.

Hyperedges Construction In this paragraph, we elaborate on how the CONSTRUCTOR identifies the hyperedges, i.e. $H = \mathcal{F}_c(X)$.

Suppose the number of hyperedges is K, which is a predefined hyperparameter. We assign a head to each hyperedge. Each head is responsible for constructing a hyperedge by selecting nodes belonging to that hyperedge.

Specifically, to construct the k-th hyperedge, the CONSTRUCTOR's k-th head outputs a column vector $\boldsymbol{m}^k \in \{0, 1\}^N$, where each element in the vector corresponds to a brain region,

$$\boldsymbol{m}^k = [\mathbb{1}(p^k_{\theta,1}), \mathbb{1}(p^k_{\theta,2}), \cdots, \mathbb{1}(p^k_{\theta,N})]^T \in \{0, 1\}^N, \tag{3}$$

where $p^k_{\theta,i} \in [0, 1], i = 1, 2, \cdots, N$ are learnable probabilities. $\mathbb{1} : [0, 1] \mapsto \{0, 1\}$ is an indicator function, which is defined as $\mathbb{1}(x) = 1$ if $x > 0.5$ and $\mathbb{1}(x) = 0$ if $x \leq 0.5$. And $\boldsymbol{m}^k$ is a column vector corresponding to the k-th hyperedge. Note that since there is no gradient defined for the indicator operation, we employ the stop-gradient technique (Oord et al., 2017; Bengio et al., 2013) to approximate it.

In the vector $\boldsymbol{m}^k$, 0 indicates nodes that are masked out, and 1 indicates nodes that are not masked. Nodes that are not masked are considered to form a hyperedge together. We use $\boldsymbol{h}^k$ to represent the masked version of X

$$\begin{aligned} \boldsymbol{h}^k &= \boldsymbol{m}^k \odot X \\ &= [\boldsymbol{m}^k_1 X_1, \boldsymbol{m}^k_2 X_2, \cdots, \boldsymbol{m}^k_N X_N] \in \mathbb{R}^{N \times d}, \end{aligned} \tag{4}$$

where $\odot$ is the broadcast element-wise multiplication. $\boldsymbol{m}^k_j$ is the j-th element of the vector $\boldsymbol{m}^k$.

We obtain K hyperedges for K sets of masks. We use H to denote the collection of all hyperedges.

$$H = (\boldsymbol{h}^1, \boldsymbol{h}^2, \cdots, \boldsymbol{h}^K). \tag{5}$$

Hyperedge Weighting After obtaining the structure (i.e. member nodes) of each hyperedge, the WEIGHTER will calculate each hyperedge's weight, which is supposed to indicate the importance of that hyperedge, based on the member nodes and their features, i.e. $\boldsymbol{w} = \mathcal{F}_w(H)$.

These weights are obtained by a $\mathrm{Readout}$ module, which is composed of: 1) summing over all the non-masked nodes feature-wisely; 2) dim reduction operation.

$$w^k = \mathrm{Readout}(\boldsymbol{h}^k) = \mathrm{DimReduction}(\boldsymbol{m}^{k^T} \boldsymbol{h}^k) \in \mathbb{R}, \tag{6}$$

where w^k is the weight of the k-th hyperedge and $\mathrm{DimReduction}$ is an MLP with ReLU activations, where the output dimension is 1. For all hyperedges, we obtain K hyperedges in total, $\boldsymbol{w} = [w^1, w^2, \cdots, w^K]^T \in \mathbb{R}^K$. Finally, these weights will be fed into the final linear head to predict the label of the instance,

$$\hat{Y} = \mathcal{F}_l(\boldsymbol{w}) \in \mathbb{R}. \tag{7}$$

In contrast to previous hypergraph construction methods (Jin et al., 2019; Xiao et al., 2019), which identify hyperedges by refining neighbors and simply aggregating node features, HYBRID makes these procedures learnable and thus is able to identify MIMR hyperedges in a data-driven way through expressive neural networks. The number of hyperedges K is decided according to the study in Appendix I.1.

Computational Complexity The computational complexity of our model is $O(N^2 K)$, which is just the same scale as that of MLPs even though we are addressing a more challenging task: identifying high-order relationships in an exponential space. Details of the complexity calculation can be found in Appendix C.

4.2. Optimization Framework

Since there are no existing IB frameworks that can be applied in our context, we propose a new IB framework named *multi-head drop-bottleneck* to optimize HYBRID. To adopt an information bottleneck view of HYBRID, we consider X, Y and H are random variables in the Markovian chain $X \leftrightarrow Y \leftrightarrow H$. According to our MIMR objective, we optimize

$$\arg\max I(H; Y) - \beta I(H; X), \tag{8}$$

where $I(\cdot; \cdot)$ denotes the mutual information. $I(H; Y)$ corresponds to the informativeness and $I(H; X)$ corresponds

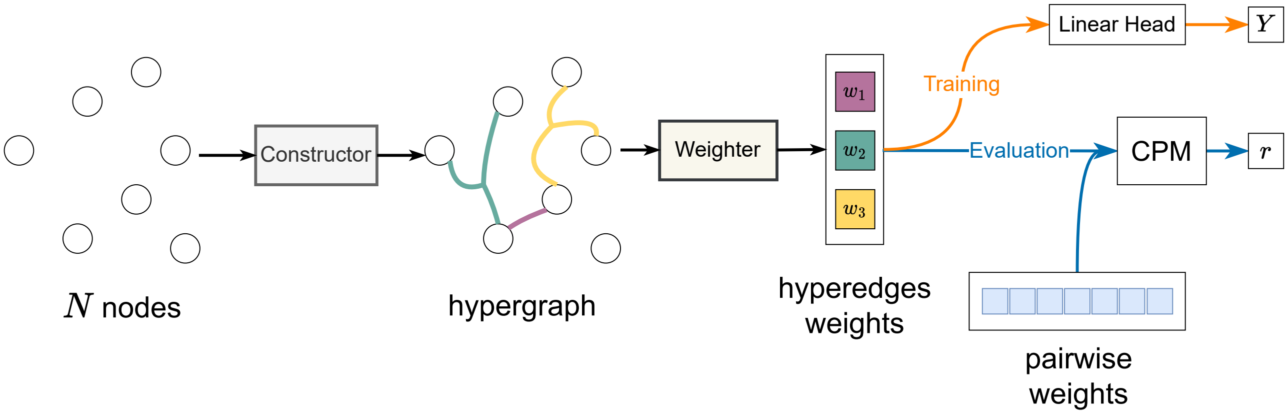


Figure 2: Overview of the HyBRiD pipeline when the total number of hyperedges $K = 3$. Hyperedge are in distinct colors for clarity. The Constructor identifies hyperedges in the hypergraph, where regions are nodes. The Weighter computes a weight for each hyperedge. These weights, representing strengths of hyperedges, are expected to be informative in terms of our target Y. There are two separate phases after obtaining weights of hyperedges: 1) Training. The model's parameters are trained under the supervision of Y; 2) Evaluation. The output weights, as well as pairwise weights, are fed into the CPM (see Appendix B).

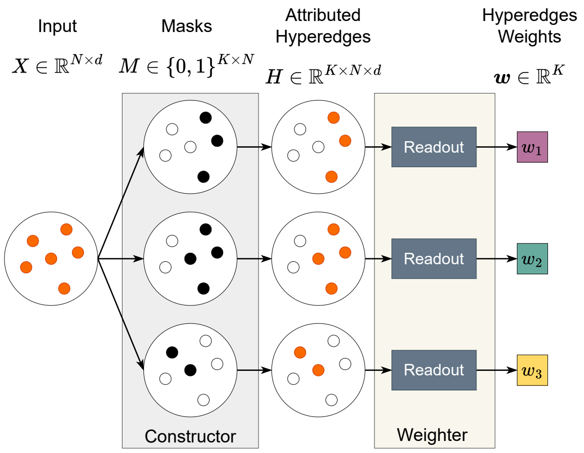

Input
Masks
Attributed Hyperedges
Hyperedges Weights
$X \in \mathbb{R}^{N\times d}$
$M \in \{0,1\}^{K\times N}$
$H \in \mathbb{R}^{K\times N\times d}$
$\boldsymbol{w} \in \mathbb{R}^{K}$
Readout
Readout
Readout
w_1
w_2
w_3
Constructor
Weighter

Figure 3: Architecture details of the Constructor and the Weighter when the number of nodes $N = 6$ and the number of hyperedges $K = 3$. At a high level, the Constructor learns the hyperedge structure by masking nodes. The Weighter computes the weight of each hyperedge based on the hyperedge's member nodes and their features.

to the redundancy. β is a coefficient trading off informativeness and redundancy. Since optimizing the mutual information for high-dimensional continuous variables is intractable, we instead optimize the lower bound of Equation 8. Specifically, for the first term (informativeness), it is easy to show

$$\begin{aligned} I(H;Y) &= \mathbb{H}[Y] - \mathbb{H}[Y|H] \\ &= \mathbb{H}[Y] + \mathbb{E}_{p(Y,H)}[\log p(Y|H)] \\ &= \mathbb{H}[Y] + \mathbb{E}_{p(Y,H)}[\log q_\phi(Y|H)] \\ &\quad + \mathbb{E}_{p(H)}[\mathrm{KL}(p(Y|H)|q_\phi(Y|H))] \\ &\geq \mathbb{H}[Y] + \mathbb{E}_{p(Y,H)}[\log q_\phi(Y|H)], \end{aligned} \tag{9}$$

where $\mathbb{H}[\cdot]$ is the entropy computation. Since there is no learnable component in the entropy of Y, we only need to optimize the second term $\mathbb{E}_{p(Y,H)}[\log q_\phi(Y|H)]$. q_ϕ can be considered as a model that predicts Y based on H, which essentially corresponds to $\mathcal{F}_l \circ \mathcal{F}_w$, where $\circ$ is the function composition. In practice, we set q_ϕ as a Gaussian model with variance 1 as most probabilistic machine learning models do for continuous data modeling (Alemi et al., 2016; Luo et al., 2019; Peng et al., 2018; Kingma & Welling, 2013).

For the second term (redundancy) in Equation 8, we have

Proposition 4.1. *(Upper bound of $I(H;X)$ in multi-head drop-bottleneck)*

$$\begin{aligned} I(H;X) \leq \sum_{k=1}^{K} I(\boldsymbol{h}^k;X) &\leq \sum_{k=1}^{K}\sum_{i=1}^{N} I(\boldsymbol{h}_i^k;X_i) \\ &= \sum_{k=1}^{K}\sum_{i=1}^{N} \mathbb{H}[X_i](1-p_{\theta,i}^k), \end{aligned} \tag{10}$$

where $\boldsymbol{h}_i^k$ and X_i is the i-th row of $\boldsymbol{h}^k$ and X respectively. $p_{\theta,i}^k$ is the mask probability in Equation 3. $\mathbb{H}$ is the entropy computation. The equality holds if and only if nodes are independent and hyperedges do not overlap. It's important to clarify that optimizing Equation 10 does not imply a penalty for overlaps. The second inequality is inspired by (Kim et al., 2021). The proof of the proposition can be found in Appendix A.

Therefore, instead of optimizing the intractable objective Equations 8, we optimize its upper bound (i.e. loss function) according to Equations 9 and 10.

$$\mathcal{L} = \|Y - \mathcal{F}_l \circ \mathcal{F}_w \circ \mathcal{F}_c(X)\|_2^2 + \beta \sum_{k=1}^{K} \sum_{i=1}^{N} \mathbb{H}[X_i](1 - p_{\theta,i}^k) \geq -I(H, Y) + \beta I(H, X). \tag{11}$$

In conclusion, The learnable components are the shallow embeddings in Equation 3, the DimReduction MLP in Equation 6 and the LinearHead $\mathcal{F}_l$ in Equation 7. For how we choose the trade-off coefficient β, see Appendix I.2 for more discussions.

5. Experiments

In this section, we conduct experiments to validate the quality of the learned hyperedges in terms of the predictive performance towards the cognition phenotype outcome. Furthermore, we conduct ablation studies to validate the key components in our model. We also analyze our results both quantitatively and qualitatively.

5.1. Predictive Performance of Hypereges

Datasets We consider two fMRI datasets:

1) *Autism Brain Imaging Data Exchange (ABIDE)* (Craddock et al., 2013) is an open-source dataset. This dataset involved resting-state fMRI of patients from 17 international sites, as well as the anatomical and phenotypic data. Regions are obtained by Craddock 200 atlas (Craddock et al., 2012). We use the preprocessed version from the official website. For prediction targets, we choose three intelligence quotients: FIQ (full-scale intelligence), VIQ (verbal intelligence quotient), and PIQ (performance intelligence).

2) *Adolescent Brain Cognitive Development (ABCD)* (Casey et al., 2018) is one of the largest public fMRI datasets. Access is limited and requires adherence to a rigorous data request procedure to acquire the data. The data is collected from $11,875$ children aged between 9 to 10 years old. The functional MRI (fMRI) data is collected from children when they were resting and when they performed three tasks (SST, EN-back, MID). We use the ABCD imaging data collected from the baseline (release 2.0) as well as the 2-year follow-up (release 3.0). In conclusion, we obtain 8 sub-datasets (we refer to them as *datasets* from now on) from 2 timepoints under 4 tasks. Regions are obtained by AAL3v1 atlas (Rolls et al., 2020). For preprocess procedures of the ABCD dataset, please refer to Appendix D for more details. For the prediction target, we consider fluid intelligence as our label. Fluid intelligence reflects the general mental ability and plays a fundamental role in various cognitive functions.

For region features, consistent with previous connectivity-based methods (Li et al., 2021; Kan et al., 2022b; Ktena et al., 2018; Said et al., 2023), we use a region's Pearson correlation coefficients to all other regions as the region features. Other details of the data preprocessing and statistics of each dataset are summarized in Appendix D.

Evaluation Metric To evaluate the quality of hyperedges obtained by HyBRiD, we use CPM (Shen et al., 2017), a standard model that could evaluate the relevance between the connectivity and the prediction target, due to its high impact in the community. In the original implementation of CPM, weights of pairwise edges are obtained by Pearson correlation between nodes. These weights, as pairwise connectivity, are fed into the CPM. CPM will output a metric that measures the overall correlation between edges and the prediction target, which can be considered as a measure of edge qualities. This process is formulated as

$$r' = \mathrm{CPM}(\boldsymbol{w}_p, Y), \tag{12}$$

where $\boldsymbol{w}_p \in \mathbb{R}^{K_p}$ denotes the pairwise edge weights and K_p is the total number of pairwise edges. r' is a metric that measures the quality of weights based on positive and negative correlations to the outcome.

To evaluate the quality of the learned weights for our model, we replace the pairwise edge weights with the learned high-order weights $\boldsymbol{w}_h \in \mathbb{R}^{K_h}$, and thus adjust Equation 12 to

$$r = \mathrm{CPM}(\boldsymbol{w}_h, Y), \tag{13}$$

Comparing r to r' reflects the quality of learned weights in terms of the prediction performance since it measures the overall correlation between weights and the prediction target. In our model, $\boldsymbol{w}_h = \boldsymbol{w}$, which is the learned hyperedge weights.

Baselines We compare our method with 3 classes of baselines: 1) *standard* method, which is exactly the classical method that predicts outcomes based on pairwise edges (Shen et al., 2017; Dadi et al., 2019; Wang et al., 2021). The comparison with standard methods shows whether the high-order connectivity has its advantage over the classical pairwise one or not. 2) *hypergraph construction* methods. We consider kNN (Huang et al., 2009), l_1 hypergraph (Wang et al., 2015), and l_2 hypergraph (Jin et al., 2019). 3) *connectivity-based phenotypic prediction* methods, which are state-of-the-art predictive models based on brain connectivity. We consider BrainNetGNN (Mahmood et al., 2021),

Table 1: r values of our hyperedges compared to baselines on the ABCD dataset. Results are averaged over 10 runs. Deterministic methods do not have standard deviations.

Type	Model	SST 1	EN-back 1	MID 1	Rest 1	SST 2	EN-back 2	MID 2	Rest 2
Standard	pairwise	0.113	0.218	0.099	0.164	0.201	0.322	0.299	0.289
Hypergraph Construction	kNN	0.115	0.268	0.168	0.127	0.257	0.266	0.238	0.315
	l_1 hypergraph	0.099	0.223	0.125	0.126	0.145	0.295	0.242	0.259
	l_2 hypergraph	$0.096_{\pm 0.002}$	$0.197_{\pm 0.003}$	$0.118_{\pm 0.003}$	$0.157_{\pm 0.016}$	$0.203_{\pm 0.005}$	$0.272_{\pm 0.004}$	$0.289_{\pm 0.011}$	$0.307_{\pm 0.006}$
Connectivity based Prediction	BrainNetGNN	$0.227_{\pm 0.060}$	$0.287_{\pm 0.043}$	$0.266_{\pm 0.046}$	$0.221_{\pm 0.040}$	$0.468_{\pm 0.058}$	$0.480_{\pm 0.068}$	$0.506_{\pm 0.057}$	$0.453_{\pm 0.028}$
	BrainGB	$0.190_{\pm 0.073}$	$0.214_{\pm 0.051}$	$0.265_{\pm 0.048}$	$0.176_{\pm 0.066}$	$0.447_{\pm 0.089}$	$0.483_{\pm 0.077}$	$0.458_{\pm 0.064}$	$0.432_{\pm 0.076}$
	BrainGNN	$0.262_{\pm 0.030}$	$0.235_{\pm 0.032}$	$0.260_{\pm 0.049}$	$0.185_{\pm 0.058}$	$0.455_{\pm 0.028}$	$0.391_{\pm 0.077}$	$0.445_{\pm 0.078}$	$0.368_{\pm 0.041}$
	BrainNetTF	$0.327_{\pm 0.084}$	$0.338_{\pm 0.056}$	$0.370_{\pm 0.098}$	$\mathbf{0.334_{\pm 0.084}}$	$0.633_{\pm 0.178}$	$0.631_{\pm 0.142}$	$0.629_{\pm 0.123}$	$0.588_{\pm 0.138}$
Ours	HyBRiD	$\mathbf{0.361_{\pm 0.058}}$	$\mathbf{0.348_{\pm 0.061}}$	$\mathbf{0.386_{\pm 0.060}}$	$0.223_{\pm 0.056}$	$\mathbf{0.738_{\pm 0.054}}$	$\mathbf{0.714_{\pm 0.037}}$	$\mathbf{0.816_{\pm 0.053}}$	$\mathbf{0.730_{\pm 0.049}}$

Table 2: r values of our hyperedges compared to baselines on the ABIDE dataset. Results are averaged over 10 runs. Deterministic methods do not have standard deviations.

Type	Model	FIQ	VIQ	PIQ
Standard	pairwise	0.052	0.124	0.056
Hypergraph Construction	kNN	0.023	0.093	0.056
	l_1 hypergraph	0.043	0.125	0.061
	l_2 hypergraph	$0.148_{\pm 0.000}$	$0.141_{\pm 0.014}$	$0.063_{\pm 0.004}$
Connectivity based Prediction	BrainNetGNN	$0.162_{\pm 0.042}$	$0.199_{\pm 0.042}$	$0.223_{\pm 0.025}$
	BrainGB	$0.125_{\pm 0.119}$	$0.154_{\pm 0.068}$	$0.157_{\pm 0.053}$
	BrainGNN	$0.105_{\pm 0.041}$	$0.176_{\pm 0.049}$	$0.159_{\pm 0.051}$
	BrainNetTF	$0.132_{\pm 0.111}$	$0.176_{\pm 0.053}$	$0.180_{\pm 0.054}$
Ours	HyBRiD	$\mathbf{0.181_{\pm 0.040}}$	$\mathbf{0.204_{\pm 0.031}}$	$\mathbf{0.245_{\pm 0.042}}$

BrainGNN (Li et al., 2021), and BrainNetTF (Kan et al., 2022b). BrainGB (Cui et al., 2022a) is a study of different brain graph neural network designs and we include its best design as a baseline. Since none of these models are able to identify hyperedge structures of brain regions, we input their last layer embeddings (each entry as a weight) into the CPM model. Note that our weights $\boldsymbol{w}$ are also last layer embeddings in HyBRiD.

Implementation & Training Details Hyperparameter choices and other details can be found in Appendix E.

Results We report r values by CPM in Table 2 and Table 1. As we can see, on the ABIDE dataset, HyBRiD consistently outperforms all the baselines on different targets, with an average improvement of 8.8% compared to the state-of-the-art model. On the ABCD dataset, HyBRiD outperforms the state-of-the-art predictive models on 7 datasets of ABCD, with an average improvement of 12.1%. The results demonstrate our model is able to learn informative hyperedges towards different phenotypic outcomes from fMRI data of various brain states. Rest 1 is the only dataset that our model reaches the second-best. We conduct analyses and propose the potential reasons in Appendix F.

Further, the comparison between our model and the pairwise baseline demonstrates the superiority of incorporating high-order relationships over relying solely on pairwise ones.

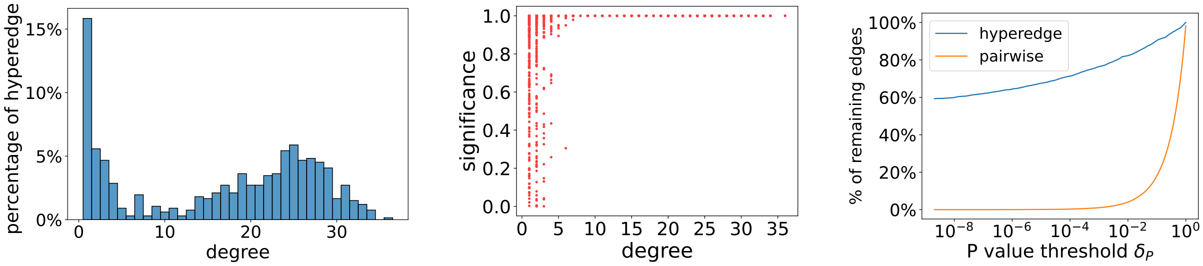
Runtime HyBRiD outperforms all other deep learning baselines in efficiency, with 87% faster than the second-fastest one (BrainNetTF). Refer to Appendix J for more runtime details.

Ablation Studies We conduct an ablation study on the effect of our masking mechanism. Specifically, we compare our model with 3 variants: 1) HyBRiD$_{\text{RndMask}}$: Replace the learnable masks with random masks with the same sparsity, initialized at the beginning of training. 2) HyBRiD$_{\text{NoMask}}$: Do not mask at all, which means all nodes and their features are visible to each head. 3) HyBRiD$_{\text{SoftMask}}$: Remove the indicator function and use $p^k_{\theta,i}$ directly in Equation 3. Ablation results are shown in Table 3. We find the original HyBRiD and the HyBRiD$_{\text{SoftMask}}$ outperform all other variants, which demonstrates the effect of learnable masks. Moreover, the original HyBRiD is better than its soft version HyBRiD$_{\text{SoftMask}}$, which demonstrates our sparse and succinct representations preserve better information than smooth ones. Other ablation studies such as the choices of the number of hyperedges and choices of β can be found in Appendix I.2.

Additional Experiments on Synthetic Dataset Since there are no ground-truth hyperedges in real-world datasets of learning informative hyperedges towards a specific outcome, we construct a synthetic dataset to verify if our model can recover the correct hyperedge structure under the MIMR objective. We use the precision, recall, and F1 score to measure the correctness of the learned hyperedges with respect to the ground truth. Although it is challenging to learn hyperedges when only supervised by the task label, we find that our model reaches high performances, with an average improvement of 28.3% in terms of the F1 score, compared to the strongest baselines. Details about the synthetic experiments can be found in Appendix G.

Table 3: Ablation studies on the masking mechanism. Results are averaged over 10 runs.

Model	SST 1	EN-back 1	MID 1	Rest 1	SST 2	EN-back 2	MID 2	Rest 2
HYBRID	$\mathbf{0.361}_{\pm \mathbf{0.058}}$	$\mathbf{0.348}_{\pm \mathbf{0.061}}$	$\mathbf{0.386}_{\pm \mathbf{0.060}}$	$\underline{0.223_{\pm 0.056}}$	$\mathbf{0.738}_{\pm \mathbf{0.054}}$	$\mathbf{0.714}_{\pm \mathbf{0.037}}$	$\mathbf{0.816}_{\pm \mathbf{0.053}}$	$\mathbf{0.730}_{\pm \mathbf{0.049}}$
HYBRID$_{\text{NoMask}}$	$0.297_{\pm 0.035}$	$0.274_{\pm 0.057}$	$\underline{0.323_{\pm 0.059}}$	$0.221_{\pm 0.034}$	$0.653_{\pm 0.036}$	$0.599_{\pm 0.059}$	$0.757_{\pm 0.021}$	$0.543_{\pm 0.038}$
HYBRID$_{\text{RndMask}}$	$0.256_{\pm 0.069}$	$0.191_{\pm 0.046}$	$0.255_{\pm 0.080}$	$0.190_{\pm 0.051}$	$0.541_{\pm 0.069}$	$0.514_{\pm 0.038}$	$0.598_{\pm 0.064}$	$0.482_{\pm 0.083}$
HYBRID$_{\text{SoftMask}}$	$\underline{0.343_{\pm 0.042}}$	$\underline{0.314_{\pm 0.040}}$	$0.320_{\pm 0.055}$	$\mathbf{0.245}_{\pm \mathbf{0.061}}$	$\underline{0.707_{\pm 0.042}}$	$\underline{0.662_{\pm 0.058}}$	$\underline{0.796_{\pm 0.031}}$	$\underline{0.655_{\pm 0.030}}$



(a) (b) (c)

Figure 4: Hyperedge profiles. **(a)** Hyperedge degree distribution of learned hyperedges. **(b)** Correlation between hyperedge degree and significance. **(c)** Comparison between the number of hyperedges and pairwise edges under different significance thresholds. The total number of hyperedges is 32. And the total number of pairwise edges is $26,896$.

Additional Experiments on Model Fit We further discuss the model's goodness of fit in Appendix H, with Mean Squared Error (MSE) as the evaluation metric. Our model outperforms the state-of-the-art model in 9 out of 11 datasets, with an average improvement of 11.9%.

5.2. Further Analysis

In this subsection, we analyze the results of our model. We mainly use the ABCD dataset in the analysis since it is much larger than the ABIDE dataset.

Hyperedge Degree Distribution We plot the hyperedge degree distribution in Figure 4a. We find there are two distinct clusters in the figure. The first cluster is hyperedges with degree ≤ 5. 1-degree and 2-degree hyperedges are special cases of our method: 1-degree hyperedges are individual nodes, which imply the contribution of individual regions to the cognition. 2-degree hyperedges reveal the importance of traditional pairwise connectivity. The other cluster concentrates around degree 25, which implies the importance of relationships of multiple regions.

Hyperedges with Higher Degree are More Significant Since CPM conducts a significance test (details can be found in Appendix B) on pairwise edges and hyperedges internally based on a linear regression model, we can obtain a P-value for each hyperedge from the significance test. We define the significance of a hyperedge as $1 - P_v \in [0, 1]$ where P_v is the P-value of that hyperedge.

The relationship between hyperedge degree and its significance is shown in Figure 4b. In this figure, we find a strong positive correlation between a hyperedge's degree and its significance, which indicates that interactions of multiple brain regions play more important roles in cognition than pairwise or individual ones. It is also worth mentioning that there is a turning point around degree 5, which corresponds to the valley around 5 in Figure 4a.

High-order relationships are Better than Pairwise Ones To compare the significance in cognition between pairwise edges and learned hyperedges, we plot the number of remaining edges under different thresholds in Figure 4c. We find out that the learned hyperedges are much more significant than pairwise ones. Also note that even if we set the threshold to an extremely strict value (1×10^{-8}), there are still 60% hyperedges considered significant. This evidence shows that our high-order relationships are much more significant than the traditional pairwise connectivity, which implies relationships involving multiple brain regions could be much more essential in cognition.

Hyperedge Case Study We visualize the most significant hyperedge of the EN-back task in Figure 5. We observe a coordinated interaction of numerous brain regions, each fulfilling specific roles. Notably, some of these regions serve multi-functional purposes:

- **Memory Processing** *ParaHippocampal_L*, *Temporal Mid*: Essential for memory encoding and retrieval,

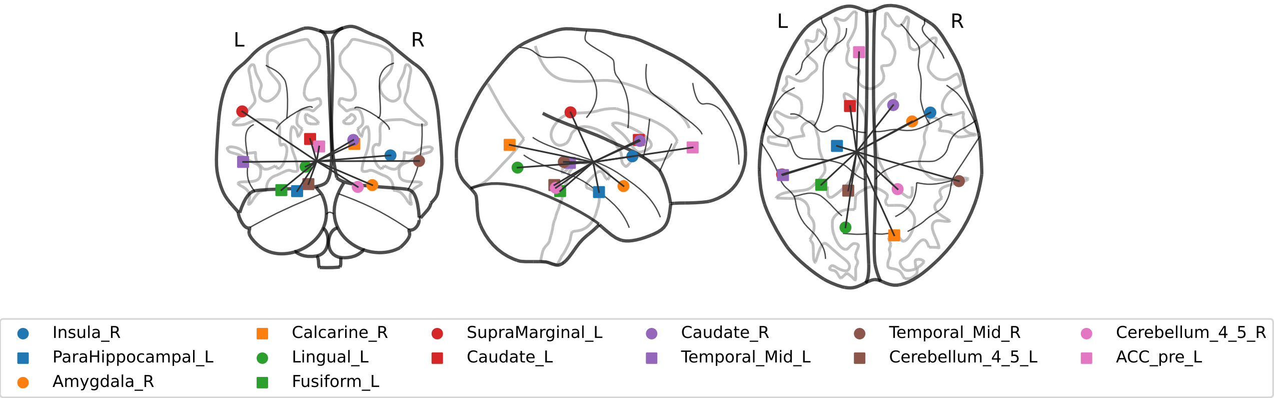


Figure 5: Visualization of the most significant hyperedge of the EN-back task.

these regions are integral to the EN-back task, facilitating the recall of previously viewed images.

- **Emotional Processing** *Amygdala_R*: The amygdala is crucial for the processing of emotions, such as fear and pleasure. Since the EN-back task involves emotional stimuli, it is reasonable that the region is connected by the hyperedge.

- **Visual Processing**: *Calcarine_R*, *Lingual_L*, *Fusiform_L*. These regions are responsible for visual perception and some of them are related to complex visual contents like symbols and human faces, which were presented during the fMRI task.

- **Sensory** *SupraMarginal_L*: It is responsible for interpreting tactile sensors and perceiving limbs location. Its involvement is likely due to the requirement for participants to engage in specific physical actions, such as pressing buttons, during the task. *Temporal Mid*: It functions in multi-modal sensory integration.

- **Motor Control** *Cerebellum*: It is primarily responsible for muscle control. *Caudate*: It plays a crucial role in motor processes. Its involvement is likely attributed to participants engaging in physical actions, like pressing buttons.

- **Cognitive Control** *ACC_pre_L*: In the EN-back task, this region is likely crucial for maintaining focus, error detection and correction, conflict management in working memory, and modulating emotional responses to the task's demands.

More visualizations about individual region importance can be found in Appendix K.

6. Conclusion

In this work, we proposed HyBRiD for identifying maximally informative yet minimally redundant (MIMR) high-order relationships of brain regions. To effectively optimize our model, we further proposed a novel information bottleneck framework and derived its theory. Our method outperforms state-of-the-art models. The result analysis shows the effectiveness of our model. We expect such advancements could benefit clinical studies, providing insights into neurological disorders, and offering improved diagnostic tools in neurology and other related fields.

Limitations HyBRiD only considers static high-order relations. Given that fMRI tasks are dynamic, including temporal changes and interactions, it will be interesting to study the evolution of these high-order relationships.

Additionally, HyBRiD does not offer a method for interpreting complex high-order relationships. This limitation is not specific to HyBRiD, but is a common challenge in analyzing such relationships.We propose a hierarchical strategy that has the potential to interpret them to a certain extent, which is detailed in Appendix L

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Ahmedt-Aristizabal, D., Armin, M. A., Denman, S., Fookes, C., and Petersson, L. Graph-based deep learning for medical diagnosis and analysis: past, present and future. *Sensors*, 21(14):4758, 2021.

Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Bai, S., Zhang, F., and Torr, P. H. Hypergraph convolution and hypergraph attention. *Pattern Recognition*, 110: 107637, 2021.

Bannadabhavi, A., Lee, S., Deng, W., Ying, R., and Li, X. Community-aware transformer for autism prediction in fmri connectome. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 287–297. Springer, 2023.

Barrasso-Catanzaro, C. and Eslinger, P. J. Neurobiological bases of executive function and social-emotional development: Typical and atypical brain changes. *Family Relations*, 65(1):108–119, 2016.

Battiston, F., Cencetti, G., Iacopini, I., Latora, V., Lucas, M., Patania, A., Young, J.-G., and Petri, G. Networks beyond pairwise interactions: Structure and dynamics. *Physics Reports*, 874:1–92, 2020.

Battiston, F., Amico, E., Barrat, A., Bianconi, G., Ferraz de Arruda, G., Franceschiello, B., Iacopini, I., Kéfi, S., Latora, V., Moreno, Y., et al. The physics of higher-order interactions in complex systems. *Nature Physics*, 17(10): 1093–1098, 2021.

Bengio, Y., Léonard, N., and Courville, A. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Bick, C., Gross, E., Harrington, H. A., and Schaub, M. T. What are higher-order networks? *SIAM Review*, 65(3): 686–731, 2023.

Boyle, R., Connaughton, M., McGlinchey, E., Knight, S. P., De Looze, C., Carey, D., Stern, Y., Robertson, I. H., Kenny, R. A., and Whelan, R. Connectome-based predictive modelling of cognitive reserve using task-based functional connectivity. *European Journal of Neuroscience*, 57(3):490–510, 2023.

Casey, B. J., Cannonier, T., Conley, M. I., Cohen, A. O., Barch, D. M., Heitzeg, M. M., Soules, M. E., Teslovich, T., Dellarco, D. V., Garavan, H., et al. The adolescent brain cognitive development (ABCD) study: Imaging acquisition across 21 sites. *Developmental cognitive neuroscience*, 32:43–54, 2018.

Chen, J. and Ying, R. Tempme: Towards the explainability of temporal graph neural networks via motif discovery. *arXiv preprint arXiv:2310.19324*, 2023.

Craddock, C., Benhajali, Y., Chu, C., Chouinard, F., Evans, A., Jakab, A., Khundrakpam, B. S., Lewis, J. D., Li, Q., Milham, M., et al. The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. *Frontiers in Neuroinformatics*, 7 (27):5, 2013.

Craddock, R. C., James, G. A., Holtzheimer III, P. E., Hu, X. P., and Mayberg, H. S. A whole brain fmri atlas generated via spatially constrained spectral clustering. *Human brain mapping*, 33(8):1914–1928, 2012.

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

Cui, H., Dai, W., Zhu, Y., Li, X., He, L., and Yang, C. Interpretable graph neural networks for connectome-based brain disorder analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 375–385. Springer, 2022b.

Dadi, K., Rahim, M., Abraham, A., Chyzhyk, D., Milham, M., Thirion, B., and Varoquaux, G. Benchmarking functional connectome-based predictive models for resting-state fMRI. *NeuroImage*, 192:115–134, 2019. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2019.02.062. URL https://www.sciencedirect.com/science/article/pii/S1053811919301594.

Do, M. T., Yoon, S.-e., Hooi, B., and Shin, K. Structural patterns and generative models of real-world hypergraphs. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 176–186, 2020.

Dubois, J., Galdi, P., Han, Y., Paul, L. K., and Adolphs, R. Resting-state functional brain connectivity best predicts the personality dimension of openness to experience. *Personality neuroscience*, 1:e6, 2018.

Finn, E. S., Shen, X., Scheinost, D., Rosenberg, M. D., Huang, J., Chun, M. M., Papademetris, X., and Constable, R. T. Functional connectome fingerprinting: Identifying individuals using patterns of brain connectivity. *Nature neuroscience*, 18(11):1664–1671, 2015.

Gao, S., Greene, A. S., Constable, R. T., and Scheinost, D. Combining multiple connectomes improves predictive modeling of phenotypic measures. *Neuroimage*, 201: 116038, 2019.

Goyal, A., Islam, R., Strouse, D., Ahmed, Z., Botvinick, M., Larochelle, H., Bengio, Y., and Levine, S. Infobot: Transfer and exploration via the information bottleneck. *arXiv preprint arXiv:1901.10902*, 2019.

Greene, A. S., Gao, S., Scheinost, D., and Constable, R. T. Task-induced brain state manipulation improves prediction of individual traits. *Nature communications*, 9(1): 2807, 2018.

Horien, C., Shen, X., Scheinost, D., and Constable, R. T. The individual functional connectome is unique and stable over months to years. *Neuroimage*, 189:676–687, 2019.

Huang, S., Elhoseiny, M., Elgammal, A., and Yang, D. Learning hypergraph-regularized attribute predictors, 2015.

Huang, Y., Liu, Q., and Metaxas, D. Video object segmentation by hypergraph cut. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1738–1745, 2009. doi: 10.1109/CVPR.2009.5206795.

Igl, M., Ciosek, K., Li, Y., Tschiatschek, S., Zhang, C., Devlin, S., and Hofmann, K. Generalization in reinforcement learning with selective noise injection and information bottleneck. *Advances in neural information processing systems*, 32, 2019.

Jin, T., Yu, Z., Gao, Y., Gao, S., Sun, X., and Li, C. Robust l_2-hypergraph and its applications. *Information Sciences*, 501:708–723, 2019.

Joshi, A., Scheinost, D., Okuda, H., Belhachemi, D., Murphy, I., Staib, L. H., and Papademetris, X. Unified framework for development, deployment and robust testing of neuroimaging algorithms. *Neuroinformatics*, 9:69–84, 2011.

Joslyn, C., Aksoy, S., Arendt, D., Jenkins, L., Praggastis, B., Purvine, E., and Zalewski, M. High performance hypergraph analytics of domain name system relationships. In *HICSS 2019 symposium on cybersecurity big data analytics*, 2019.

Kan, X., Cui, H., Lukemire, J., Guo, Y., and Yang, C. Fbnetgen: Task-aware GNN-based fMRI analysis via functional brain network generation. In *International Conference on Medical Imaging with Deep Learning*, pp. 618–637. PMLR, 2022a.

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

Kim, J., Kim, M., Woo, D., and Kim, G. Drop-Bottleneck: Learning discrete compressed representation for noise-robust exploration. *arXiv preprint arXiv:2103.12300*, 2021.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Knauff, M. and Wolf, A. G. Complex cognition: the science of human reasoning, problem-solving, and decision-making, 2010.

Ktena, S. I., Parisot, S., Ferrante, E., Rajchl, M., Lee, M., Glocker, B., and Rueckert, D. Metric learning with spectral graph convolutions on brain connectivity networks. *NeuroImage*, 169:431–442, 2018.

Kucian, K., Loenneker, T., Dietrich, T., Dosch, M., Martin, E., and Von Aster, M. Impaired neural networks for approximate calculation in dyscalculic children: a functional MRI study. *Behavioral and Brain Functions*, 2(1): 1–17, 2006.

Kucian, K., von Aster, M., Loenneker, T., Dietrich, T., and Martin, E. Development of neural networks for exact and approximate calculation: A fMRI study. *Developmental neuropsychology*, 33(4):447–473, 2008.

Li, H.-J., Hou, X.-H., Liu, H.-H., Yue, C.-L., He, Y., and Zuo, X.-N. Toward systems neuroscience in mild cognitive impairment and Alzheimer's disease: A meta-analysis of 75 fMRI studies. *Human brain mapping*, 36(3):1217–1232, 2015a.

Li, H.-J., Hou, X.-H., Liu, H.-H., Yue, C.-L., Lu, G.-M., and Zuo, X.-N. Putting age-related task activation into large-scale brain networks: a meta-analysis of 114 fMRI studies on healthy aging. *Neuroscience & Biobehavioral Reviews*, 57:156–174, 2015b.

Li, X., Dvornek, N. C., Zhou, Y., Zhuang, J., Ventola, P., and Duncan, J. S. Graph neural network for interpreting task-fMRI biomarkers. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part V 22*, pp. 485–493. Springer, 2019.

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

Li, Y., Li, Q., Li, T., Zhou, Z., Xu, Y., Yang, Y., Chen, J., and Guo, H. Construction and multiple feature classification based on a high-order functional hypernetwork on fMRI data. *Frontiers in Neuroscience*, 16:848363, 2022.

Liu, Q., Sun, Y., Wang, C., Liu, T., and Tao, D. Elastic net hypergraph learning for image clustering and semi-supervised classification. *IEEE Transactions on Image Processing*, 26(1):452–463, 2017. doi: 10.1109/TIP.2016.2621671.

Logue, S. F. and Gould, T. J. The neural and genetic basis of executive function: attention, cognitive flexibility, and response inhibition. *Pharmacology Biochemistry and Behavior*, 123:45–54, 2014.

Luo, Y., Liu, P., Guan, T., Yu, J., and Yang, Y. Significance-aware information bottleneck for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6778–6787, 2019.

Mahmood, U., Fu, Z., Calhoun, V. D., and Plis, S. A deep learning model for data-driven discovery of functional connectivity. *Algorithms*, 14(3):75, 2021.

Miao, S., Liu, M., and Li, P. Interpretable and generalizable graph learning via stochastic attention mechanism. In *International Conference on Machine Learning*, pp. 15524–15543. PMLR, 2022a.

Miao, S., Luo, Y., Liu, M., and Li, P. Interpretable geometric deep learning via learnable randomness injection. *arXiv preprint arXiv:2210.16966*, 2022b.

Oord, A. v. d., Vinyals, O., and Kavukcuoglu, K. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*, 2017.

Peng, R., Duan, Q., Wang, H., Ma, J., Jiang, Y., Tu, Y., Jiang, X., and Zhao, J. Came: Contrastive automated model evaluation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20121–20132, 2023.

Peng, R., Zou, H., Wang, H., Zeng, Y., Huang, Z., and Zhao, J. Energy-based automated model evaluation. *arXiv preprint arXiv:2401.12689*, 2024.

Peng, X. B., Kanazawa, A., Toyer, S., Abbeel, P., and Levine, S. Variational discriminator bottleneck: Improving imitation learning, inverse RL, and GANs by constraining information flow. *arXiv preprint arXiv:1810.00821*, 2018.

Pornpattananangkul, N., Hariri, A. R., Harada, T., Mano, Y., Komeda, H., Parrish, T. B., Sadato, N., Iidaka, T., and Chiao, J. Y. Cultural influences on neural basis of inhibitory control. *NeuroImage*, 139:114–126, 2016.

Reineberg, A. E., Banich, M. T., Wager, T. D., and Friedman, N. P. Context-specific activations are a hallmark of the neural basis of individual differences in general executive function. *NeuroImage*, 249:118845, 2022.

Rolls, E. T., Huang, C.-C., Lin, C.-P., Feng, J., and Joliot, M. Automated anatomical labelling atlas 3. *Neuroimage*, 206:116189, 2020.

Rosas, F. E., Mediano, P. A., Gastpar, M., and Jensen, H. J. Quantifying high-order interdependencies via multivariate extensions of the mutual information. *Physical Review E*, 100(3):032305, 2019.

Rosenberg, M. D., Finn, E. S., Constable, R. T., and Chun, M. M. Predicting moment-to-moment attentional state. *Neuroimage*, 114:249–256, 2015.

Rosenberg, M. D., Finn, E. S., Scheinost, D., Papademetris, X., Shen, X., Constable, R. T., and Chun, M. M. A neuromarker of sustained attention from whole-brain functional connectivity. *Nature neuroscience*, 19(1):165–171, 2016.

Rosenberg, M. D., Scheinost, D., Greene, A. S., Avery, E. W., Kwon, Y. H., Finn, E. S., Ramani, R., Qiu, M., Constable, R. T., and Chun, M. M. Functional connectivity predicts changes in attention observed across minutes, days, and months. *Proceedings of the National Academy of Sciences*, 117(7):3797–3807, 2020.

Said, A., Bayrak, R. G., Derr, T., Shabbir, M., Moyer, D., Chang, C., and Koutsoukos, X. Neurograph: Benchmarks for graph machine learning in brain connectomics. *arXiv preprint arXiv:2306.06202*, 2023.

Santoro, A., Battiston, F., Petri, G., and Amico, E. Higher-order organization of multivariate time series. *Nature Physics*, 19(2):221–229, 2023.

Satterthwaite, T. D., Wolf, D. H., Roalf, D. R., Ruparel, K., Erus, G., Vandekar, S., Gennatas, E. D., Elliott, M. A., Smith, A., Hakonarson, H., et al. Linked sex differences in cognition and functional connectivity in youth. *Cerebral cortex*, 25(9):2383–2394, 2015.

Semedo, J. D., Zandvakili, A., Machens, C. K., Byron, M. Y., and Kohn, A. Cortical areas interact through a communication subspace. *Neuron*, 102(1):249–259, 2019.

Shen, X., Papademetris, X., and Constable, R. T. Graph-theory based parcellation of functional subunits in the brain from resting-state fmri data. *Neuroimage*, 50(3): 1027–1035, 2010.

Shen, X., Tokoglu, F., Papademetris, X., and Constable, R. Groupwise whole-brain parcellation from resting-state fMRI data for network node identification. *NeuroImage*, 82:403–415, November 2013. ISSN 10538119. doi: 10.1016/j.neuroimage.2013.05.081.

Shen, X., Finn, E. S., Scheinost, D., Rosenberg, M. D., Chun, M. M., Papademetris, X., and Constable, R. T. Using connectome-based predictive modeling to predict individual behavior from brain connectivity. *nature protocols*, 12(3):506–518, 2017.

Thomas, A., Ré, C., and Poldrack, R. Self-supervised learning of brain dynamics from broad neuroimaging data. *Advances in Neural Information Processing Systems*, 35: 21255–21269, 2022.

Tishby, N., Pereira, F. C., and Bialek, W. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Varley, T. F., Pope, M., Faskowitz, J., and Sporns, O. Multivariate information theory uncovers synergistic subsystems of the human cerebral cortex. *Communications biology*, 6(1):451, 2023.

Wang, J., Chen, J., and Su, B. Toward auto-evaluation with confidence-based category relation-aware regression. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023a.

Wang, M., Liu, X., and Wu, X. Visual classification by l_1-hypergraph modeling. *IEEE Transactions on Knowledge and Data Engineering*, 27(9):2564–2574, 2015. doi: 10.1109/TKDE.2015.2415497.

Wang, R., He, X., Yu, R., Qiu, W., An, B., and Rabinovich, Z. Learning efficient multi-agent communication: An information bottleneck approach. In *International Conference on Machine Learning*, pp. 9908–9918. PMLR, 2020.

Wang, S., Liu, Y., Xu, W., Tian, X., and Zhao, Y. Inference-based statistical network analysis uncovers star-like brain functional architectures for internalizing psychopathology in children. *arXiv preprint arXiv:2309.11349*, 2023b.

Wang, Y., Kang, J., Kemmer, P. B., and Guo, Y. An efficient and reliable statistical method for estimating functional connectivity in large scale brain networks using partial correlation. *Frontiers in neuroscience*, 10:123, 2016.

Wang, Z., Goerlich, K. S., Ai, H., Aleman, A., Luo, Y.-j., and Xu, P. Connectome-Based Predictive Modeling of Individual Anxiety. *Cerebral Cortex*, 31(6): 3006–3020, 01 2021. ISSN 1047-3211. doi: 10.1093/cercor/bhaa407. URL https://doi.org/10.1093/cercor/bhaa407.

Wu, T., Ren, H., Li, P., and Leskovec, J. Graph information bottleneck. *Advances in Neural Information Processing Systems*, 33:20437–20448, 2020.

Xiao, L., Wang, J., Kassani, P. H., Zhang, Y., Bai, Y., Stephen, J. M., Wilson, T. W., Calhoun, V. D., and Wang, Y.-P. Multi-hypergraph learning-based brain functional connectivity analysis in fMRI data. *IEEE transactions on medical imaging*, 39(5):1746–1758, 2019.

Xu, D., Cheng, W., Luo, D., Chen, H., and Zhang, X. InfoGCL: Information-aware graph contrastive learning. *Advances in Neural Information Processing Systems*, 34: 30414–30425, 2021.

Ying, Z., You, J., Morris, C., Ren, X., Hamilton, W., and Leskovec, J. Hierarchical graph representation learning with differentiable pooling. *Advances in neural information processing systems*, 31, 2018.

Yu, J., Xu, T., Rong, Y., Bian, Y., Huang, J., and He, R. Graph information bottleneck for subgraph recognition. *arXiv preprint arXiv:2010.05563*, 2020.

Yu, J., Cao, J., and He, R. Improving subgraph recognition with variational graph information bottleneck. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19396–19405, 2022.

Zhang, Z., Lin, H., Gao, Y., and BNRist, K. Dynamic hypergraph structure learning. In *IJCAI*, pp. 3162–3169, 2018.

Zhang, Z., Feng, Y., Ying, S., and Gao, Y. Deep hypergraph structure learning, 2022.

Zu, C., Gao, Y., Munsell, B., Kim, M., Peng, Z., Zhu, Y., Gao, W., Zhang, D., Shen, D., and Wu, G. Identifying high order brain connectome biomarkers via learning on hypergraph. In *Machine Learning in Medical Imaging: 7th International Workshop, MLMI 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 17, 2016, Proceedings 7*, pp. 1–9. Springer, 2016.

A. Proof of the Upper Bound

In this section, we prove the upper bound of $I(H;X)$ in *multi-head drop-bottleneck* in Equation 10.

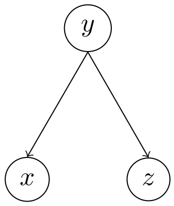


Figure 6: The graphical model of random variables X, Y and Z.

Lemma A.1. *Given random variables X, Y and Z. Their relationships are described in the graphical model illustrated in Figure 6. We have*

$$I(X;Y|Z) \leq I(X;Y) \tag{14}$$

Proof.

$$\begin{aligned}
&I(X;Y|Z) - I(X;Y) \\
&\quad= \int p(x,y,z)\log\frac{p(z)p(x,y,z)}{p(x,z)p(y,z)}\mathrm{d}x\mathrm{d}z\mathrm{d}y \\
&\quad- \int p(x,y)\log\frac{p(x,y)}{p(x)p(y)}\mathrm{d}x\mathrm{d}y \\
&\quad= \int p(x,y,z)\log\frac{p(z)p(x,y,z)}{p(x,z)p(y,z)}\mathrm{d}x\mathrm{d}z\mathrm{d}y \\
&\quad- \int p(x,y,z)\log\frac{p(x,y)}{p(x)p(y)}\mathrm{d}x\mathrm{d}y\mathrm{d}z \\
&\quad= \int p(x,y,z)\log\frac{p(z)p(x,y,z)p(x)p(y)}{p(x,z)p(y,z)p(x,y)}\mathrm{d}x\mathrm{d}z\mathrm{d}y \\
&\quad= \int p(x,y,z)\log\frac{p(z)p(x,z|y)p(y)p(x)p(y)}{p(x,z)p(y,z)p(x,y)}\mathrm{d}x\mathrm{d}z\mathrm{d}y \\
&\quad= \int p(x,y,z)\log\frac{p(z)p(x|y)p(z|y)p(y)p(x)p(y)}{p(x,z)p(y,z)p(x,y)}\mathrm{d}x\mathrm{d}z\mathrm{d}y \\
&\quad= \int p(x,y,z)\log\frac{p(x)p(z)}{p(x,z)}\mathrm{d}x\mathrm{d}z\mathrm{d}y \\
&\quad= \int p(x,z)\log\frac{p(x)p(z)}{p(x,z)}\mathrm{d}x\mathrm{d}z \\
&\quad= -I(X;Z) \leq 0
\end{aligned} \tag{15}$$

which finishes the proof.

□

Corollary A.2. *Given the same graphical model 6, we have*

$$I(X,Z;Y) \leq I(X;Y) + I(Z;Y) \tag{16}$$

Proof. Using the chain rule of mutual information, we obtain

$$I(X,Z;Y) = I(X;Y|Z) + I(Z;Y) \tag{17}$$

According to Lemma A.1, we have $I(X;Y|Z) \leq I(X;Y)$, which finishes the proof.

□

Theorem A.3. *For random variables in Equation 10, we have*

$$I(H;X) \leq \sum_{k=1}^{K} I(\boldsymbol{h}^k;X) \tag{18}$$

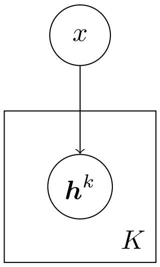


Figure 7: The graphical model of random variables $\boldsymbol{h}^k$ and X

Proof. According to the definitions of X and H, which are described in Section 4, we can draw a graphical model of them in Figure 7. Define a new random variable $\boldsymbol{h}^{k_1:k_2} = [\boldsymbol{h}^{k_1}, \boldsymbol{h}^{k_1+1}, \cdots, \boldsymbol{h}^{k_2-1}, \boldsymbol{h}^{k_2}]$, which is a concatenation from $\boldsymbol{h}^{k_1}$ to $\boldsymbol{h}^{k_2}$. According to Corollary A.2 we have

$$\begin{aligned} I(H;X) &\leq I(\boldsymbol{h}^1, X) + I(\boldsymbol{h}^{2:K}, X) \\ &\leq I(\boldsymbol{h}^1, X) + I(\boldsymbol{h}^2, X) + I(\boldsymbol{h}^{3:K}, X) \\ &\leq I(\boldsymbol{h}^1, X) + I(\boldsymbol{h}^2, X) + I(\boldsymbol{h}^3, X) + \cdots \\ &\leq \sum_{k=1}^{K} I(\boldsymbol{h}^k;X) \end{aligned} \tag{19}$$

□

Theorem A.4. *(proposition 1)*

$$I(H;X) \leq \sum_{k=1}^{K}\sum_{i=1}^{N} I(\boldsymbol{h}_i^k;X_i) = \sum_{k=1}^{K}\sum_{i=1}^{N} \mathbb{H}[X_i](1-p_{\theta,i}^k) \tag{20}$$

Proof. Given **Theorem 2**. It suffices to prove

$$I(\boldsymbol{h}^k;X) \leq \sum_{i=1}^{N} I(\boldsymbol{h}_i^k;X_i) = \sum_{i=1}^{N} \mathbb{H}[X_i](1-p_{\theta,i}^k), \quad \forall 1 \leq k \leq K \tag{21}$$

And this is exactly the conclusion in (Kim et al., 2021) if we consider $\boldsymbol{h}^k$ and X as X and Z respectively in their paper. □

B. Connectome-Based Predictive Modeling

Shen et al. (2017); Finn et al. (2015) have shown tremendous promise in recent years in detecting imaging biomarkers by CPM (connectome-based predictive modeling) (Rosenberg et al., 2015; Dubois et al., 2018; Rosenberg et al., 2020; 2016). Such a model, based chiefly on functional MRI data, can measure the significance of the input edge weights, which is revealed by a correlation coefficient that reflects the correlation between the edge weights and the neurological outcomes. One could expect a large correlation coefficient to indicate the high quality of edge weights. We utilize the CPM as an evaluation model to evaluate the quality of our learned hyperedges. Here is a pipeline overview of the CPM process:

1. **Connectivity Calculation**: For each subject, compute the Pearson correlation coefficients for each possible pair of brain regions. This is based on the fMRI series of those regions.

2. **Edge Significance**: Calculate the correlation between each brain connectivity edge and the outcome of interest (e.g., cognition scores) across all subjects. The correlation of an edge indicates its significance.
3. **Edge Selection**: Identify significant connectivity edges. These are the edges where the correlation values are greater than a predetermined significance threshold.
4. **Weight Summation**: For each subject, sum the weights of the significant edges identified in the previous step to derive a single summary score (scalar).
5. **Model Fitting**: Fit a linear model that predicts the neurological outcomes based on the summed weights, where each subject is a sample.
6. **Model Evaluation**: Across all subjects, calculate the correlation of predicted values and the neurological outcomes. Note that this correlation coefficient is equivalent to the one between the summed weights and the outcomes, and is exactly the metric r we use to evaluate our hyperedges in Equation 13.

Since positive edges and negative edges will cancel out with each other when being summed, we adopt the combining strategy in (Boyle et al., 2023).

CPM Measures the Quality of Edge Weights According to step 6, the evaluation of the predictive model could be measured by the correlation between predicted and ground-truth outcomes (Shen et al., 2017). Since CPM is a linear model that predicts the outcome based on the sum of significant edge weights, the correlation is equal to the correlation between the sum and ground-truth outcomes (which is exactly the r in Equation 13). Hence, one can expect a larger correlation if the edge weights are more correlated (and thus are more predictive).

Significance of Edges in CPM In step 2, CPM obtains a correlation coefficient r^k for each edge weight w^k and the cognition score Y across all subjects. Consider a classical hypothesis test $H_0 : r^k = 0, H_1 : r^k \neq 0$. Assume w^k and Y are drawn from independent normal distribution (corresponds to H_0), the probability density function of correlation coefficient r^k is

$$f(r^k) = \frac{\left(1 - {r^k}^2\right)^{n/2-2}}{\mathrm{B}\left(\frac{1}{2}, \frac{n}{2} - 1\right)},$$

where n is the number of samples and B is the beta function. Based on the distribution, we obtain a P-value for the k-th edge, which is used to measure the significance of the edge.

C. Computational Complexity

Suppose we have N regions as N nodes and K hyperedges. For the CONSTRUCTOR, the unignorable computation is from the mask operation. The computational complexity of this step for each hyperedge is $O(Nd)$ since it does a pairwise multiplication operation of a matrix of size $N \times d$. Given that there are K hyperedges, the total complexity is $O(NKd)$. For the WEIGHTER, the computation is from the dim reduction operation and the linear head. The dim reduction operation is an MLP. In this work, the hidden dimensions are a fraction of the original feature dimension. Therefore, the complexity of the dim-reduction MLP is $O(d^2)$. The linear head only contributes $O(d)$, which is neglectable. As a result, the computational complexity of the whole model is $O(NKd + d^2) = O(N^2K)$ since the feature dimension is equal to the number of regions (See Appendix D for features we used). This complexity is just at the same scale as that of MLPs even though we are addressing a more challenging task: identifying high-order relationships in an exponential space.

D. Dataset details

D.1. Preprocessing of ABCD Dataset

We use the preprocessed ABIDE dataset from the official website. We preprocessed the restricted ABCD dataset ourselves. Below is the preprocessing procedure of the ABCD dataset.

raw data to voxel-level fMRI time series The fMRI data is processed using BioImage Suite (Joshi et al., 2011). First, we performed motion correction and slice-time correction using SPM5; and via BioImage Suite, the data were registered to a standardized $3mm \times 3mm \times 3mm$ common space, where we generated masks representing white matter, gray matter, and cerebrospinal fluid (CSF) and computed the mean time courses for both white matter and CSF. We orthogonalized each

Table 4: Statistics of datasets we use. ABIDE dataset only contains resting-state fMRI data. ABCD contains resting-state and task-based data. We use ABCD of 2 timepoints. For example, *Rest 1* means resting-state fMRI data from timepoint 1.

Dataset	ABIDE			ABCD							
	FIQ	VIQ	PIQ	Rest 1	SST 1	EN-back 1	MID 1	Rest 2	SST 2	EN-back 2	MID 2
#(instances)	1035	1035	1035	1676	1673	1678	1678	1949	1053	1044	1062
length of time series	196	196	196	375	437	362	403	375	437	362	403

gray matter time course with respect to the mean time courses of both white matter and CSF, and we orthogonalized each gray matter time course to the six motion-related signals via SPM5. We then applied a bandpass Butterworth filter with a frequency range of 0.02Hz to 0.1Hz to the orthogonalized time courses. We used a Gaussian kernel with a full-width at half-maximum (FWHM) of 6mm to enhance spatial coherence and spatial smoothing. Lastly, we removed the linear trend from all signals in accordance with the methodology detailed in (Shen et al., 2013). We deleted scans with more than 0.10 mm mean frame-to-frame displacement. Additional details about the standard preprocessing procedures, such as slice time and motion correction, and registration to the MNI template can be found in (Greene et al., 2018) and (Horien et al., 2019).

voxel-level fMRI time series to region features The fMRI time series data of a human subject is represented in four dimensions (3 spatial dimensions + 1 temporal dimension), which can be imagined as a temporal sequence of 3D images. First, brain images are parceled into regions (or nodes) using the AAL3v1 atlas (Rolls et al., 2020). Following previous works (Kan et al., 2022b; Li et al., 2021; Thomas et al., 2022), each region's time series is obtained by averaging all voxels in that region. Consistent with previous connectivity-based methods (Li et al., 2021; Kan et al., 2022b; Ktena et al., 2018; Said et al., 2023), for each region, we use its Pearson correlation coefficients to all regions as its features. We randomly split the data into train, validation, and test sets in a stratified fashion. The split ratio is 8:1:1.

D.2. Dataset Statistics

The statistics of the number of instances and the time series length are summarized in Table 4.

E. Training Details

Due to the data scarcity, we found training on single sub-datasets of the ABCD dataset leads to severe overfitting. To mitigate this, we train our model as well as all the baselines on all datasets together and report the results individually. Note that on the ABIDE dataset, we train our model under the three targets separately since we don't encounter such an issue on the ABIDE dataset.

Hardware We train our model on a machine with an Intel Xeon Gold 6326 CPU and RTX A5000 GPUs.

Software See Table 5 for the software we used and the versions.

Table 5: Software versions.

software	version
python	3.8.13
pytorch	1.11.0
cudatoolkit	11.3
numpy	1.23.3
ai2-tango	1.2.0
nibabel	4.0.2

Hyperparameter Choices The hyperparameters selection is shown in Table 6. Some crucial hyperparameters ablation experiments can be found in Appendix I.

Table 6: Hyperparameter choices.

notation	meaning	value
lr	learning rate	1×10^{-3}
K	number of hyperedges	32
β	trade-off coefficients information bottleneck	0.2
$[h_1, h_2, h_3]$	hidden sizes of the dim reduction MLP	$[32, 8, 1]$
B	batch size	64

F. Why is HYBRID not the Best on the Rest 1 Dataset?

At present, we do not fully understand the reasons for the phenomena observed. The hypotheses outlined below are based on our preliminary observations and analyses.

Hypothesis 1: Sparsity in Resting-State Connectivity We propose that resting-state connectivity is sparser and more diffuse compared to task-based connectivity. This assertion is refined from our initial claim that the connectivity relevant to the phenotypic outcomes during the resting state is notably sparser than during the task-based state.

Upon examining traditional pairwise connections, we summarized the number of significant connections most related to the phenotypic outcome in Table 7. Our analysis indicates that Rest 1 exhibits the fewest phenotype-related connections, suggesting that these connections are indeed sparse. We did not evaluate high-order relationships due to the lack of ground truth data for such connections. Further investigation is required to validate this hypothesis.

	SST 1	EN-back 1	MID 1	Rest 1	SST 2	EN-back 2	MID 2	Rest 2
# (connections)	3164	3414	2989	2036	3112	3014	2692	3896

Table 7: The number of significant pairwise edges selected by CPM in different tasks and timepoints.

Hypothesis 2: Comparative Quality of Data Across Timepoints We hypothesize that the data quality at timepoint 1 is inferior to that at timepoint 2. The performance metrics, as shown in the Table 8, are better at timepoint 2 than at timepoint 1. Moreover, timepoint 1 exhibits more severe overfitting compared to timepoint 2. This discrepancy may be attributed to the higher relevance of the information (pertaining to the outcome) and lower noise levels at timepoint 2. This hypothesis also requires further verification through detailed investigations and experiments.

	SST 1	EN-back 1	MID 1	Rest 1	SST 2	EN-back 2	MID 2	Rest 2
train	0.988	0.984	0.984	0.984	0.982	0.977	0.978	0.976
test	0.361	0.348	0.386	0.223	0.738	0.714	0.816	0.730

Table 8: Performances on training and test dataset in different tasks and timepoints.

Conclusion Brain activities during resting states are not driven by external tasks, leading to more diffuse and less predictable patterns of activation. The low data quality of ABCD timepoint 1 (compared to timepoint 2) even intensifies this issue. Therefore, high-order relations might not be a good inductive bias on the Rest 1 dataset since the connections might be much more sparse and involve fewer nodes.

G. Evaluation on Synthetic Dataset

Dataset Synthesis Our synthetic dataset is constructed as follows:

1) *Structure Generation*: For each hyperedge, we randomly sample the hyperedge degree d from a discrete uniform distribution between 2 and d_{max}. After that, we randomly sample d nodes as members of this hyperedge.

Table 9: Precision, recall and F1 score on synthetic dataset under different K, where K is the number of hyperedges. The performance of the Random method increases as K increases because the Hungarian algorithm is likely to provide better matches when there are more candidates.

Metric	$K = 1$			$K = 5$			$K = 10$			$K = 30$		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Random	0.158	0.094	0.110	0.182	0.170	0.171	0.224	0.218	0.218	0.288	0.250	0.266
kNN	0.533	**1.000**	0.696	0.247	0.241	0.243	0.144	0.156	0.148	0.139	0.128	0.133
l_1 hypergraph	0.160	**1.000**	0.276	0.180	0.016	0.028	0.135	0.016	0.028	0.256	0.026	0.047
l_2 hypergraph	0.987	0.825	0.897	0.412	0.631	0.494	0.276	0.499	0.351	0.233	0.470	0.310
HGCN	0.625	**1.000**	0.769	0.196	0.477	0.278	0.188	0.379	0.251	0.277	0.364	0.314
HyBRiD (Ours)	**1.000**	**1.000**	**1.000**	**0.595**	**0.680**	**0.631**	**0.504**	**0.604**	**0.549**	**0.428**	**0.393**	**0.401**

2) *Feature generation*: For each hyperedge, we randomly sample a scalar value v from the uniform distribution $U(0, 1)$. We then randomly sample the features from $U(0, 2v)$ for each node in this hyperedge.

3) *Label Generation*: For each hyperedge, we calculate the maximum of its node features as the summary of the hyperedge. We sum the summaries of all hyperedges to get a single value Y as the label of the hypergraph.

Note that consistent with our settings of real-world datasets, the structure is shared across all hypergraphs, while node features and hyperedge weights are different.

Metrics We use the macro-precision, recall and F1 score to measure the correctness of the learned hyperedges with respect to the ground-truth ones. Note that when learning on the synthetic dataset, the order of hyperedge may differ from the ground truths. Therefore, the Hungarian algorithm is employed to match the learned hyperedges with the ground truth. The precision, recall and F1 score are calculated after matching.

Baselines We use the hyperedge construction methods in Section 5, i.e. kNN(Huang et al., 2009), l_1 hypergraph(Wang et al., 2015) and l_2 hypergraph(Jin et al., 2019) as our baselines. Besides, we implement a hypergraph structure learning model. This is a hypergraph convolutional neural network (HGCN) (Bai et al., 2021), where both the parameters of layers and the hypergraph structure (the incidence matrix) are learnable.

Implementaton Details The maximum degree d_{max} is set to 34, which is the maximum degree of hyperedges of the ABCD dataset according to the analysis in Section 5. The number of nodes is set to 164, which is the number of regions of AAL3v1 atlas, used in the ABCD experiment. We conduct the experiments under different numbers of hyperedges (i.e. $K = 1, 5, 10, 30$).

Results We report the results under different numbers of hyperedges. Although this is a hard task, our model consistently outperforms all the baselines significantly, with an average improvement of 28.3% in terms of the F1 score. This demonstrates that our model can learn the hyperedges well under the MIMR objective, without the direct supervision of hyperedge ground truths. This also inspires us to use automated model evaluation techniques (Wang et al., 2023a; Peng et al., 2023; 2024) on real-world data in the future where labelled high-order relationships are not readily accessible.

H. Model Fit Performance

We report the goodness of fit of our model and the state-of-the-art baseline in Table 10, with the Mean Square Error (MSE) as the metric.

Model	FIQ	VIQ	PIQ
BrainNetTF	5.917	5.429	3.355
HyBRiD (ours)	**3.477**	**4.331**	**2.806**

Table 10: Performance of the model and baselines (MSE) in fitting the target on the ABIDE dataset.

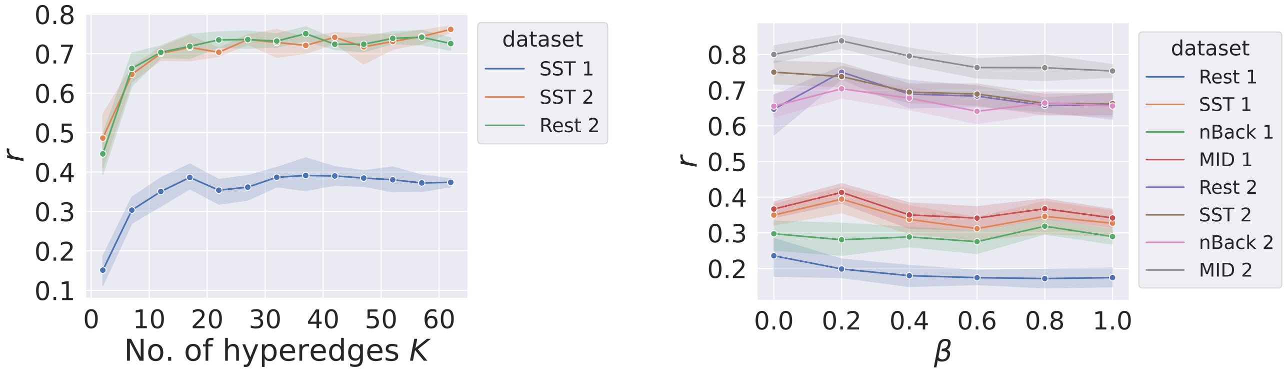
Model	SST 1	EN-back 1	MID 1	Rest 1	SST 2	EN-back 2	MID 2	Rest 2
BrainNetTF	1.153	1.373	**1.147**	**1.244**	0.616	0.677	0.720	0.750
HyBRiD (ours)	**1.031**	**1.348**	1.275	1.470	**0.604**	**0.641**	**0.543**	**0.464**

Table 11: Performance of the model and baselines (MSE) in fitting the target on the ABCD dataset.

I. More Ablation Studies

I.1. Choices of Number of Hyperedges K

As explained in Section 4, we use K heads for K hyperedges. We study the correlation between the r value and the number of hyperedges on three datasets:



(a) The performance with increasing the number of hyperedges through 2, 7, 12, 17, 22, 27, 32, 37, 52, 57, 62 on three datasets.

(b) Performance with increasing β from 0.1 to 1.0 with step 0.1 on all datasets.

Figure 8: Studies on the choice of two key hyperparameters, β and K, in our model.

From Figure 8a, we find that the overall performance increases dramatically before $K = 17$, but becomes stable and close to saturation after $K = 32$. To improve the efficiency while ensuring the performance, we choose $K = 32$.

I.2. Choices of the Trade-off Coefficient β

In our optimization objective 8, β acts as a trade-off parameter, which is a non-negative scalar that determines the weight given to the second term relative to the first. To study its fluence to the performance, we plot the model performances on all datasets under different β in Figure 8b. We can see performances on 3 datasets (Rest 1, SST 2) consistently decrease when β increases. However, on other 5 datasets (SST 1, MID 1, Rest 2, nBack 2, MID 2), we can observe a peak at $\beta = 0.2$. Accordingly, we adopt $\beta = 0.2$.

J. Runtime Comparison

Figure 9 summarizes the per-batch training time of all deep learning models. We find that HyBRiD is the most efficient one, with 87% faster than the second one (BrainNetTF) and at least 1255% faster than the GNN-based ones.

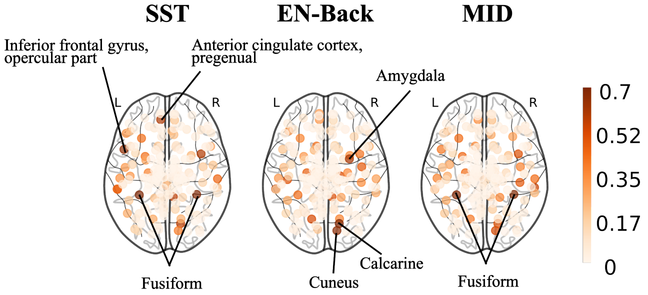

SST
EN-Back
MID
Inferior frontal gyrus, opercular part
Anterior cingulate cortex, pregenual
Amygdala
L
R
L
R
L
R
Fusiform
Cuneus
Calcarine
Fusiform
0.7
0.52
0.35
0.17
0

Figure 10: Visualization of the frequency of each region under different fMRI tasks.

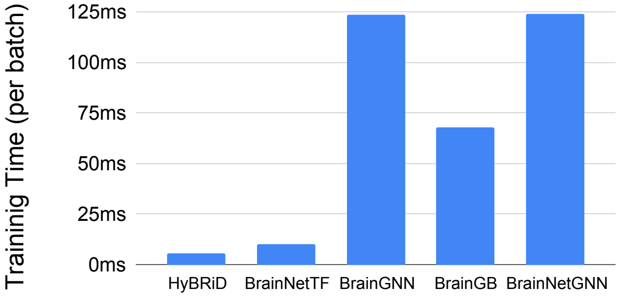


Figure 9: Training time per batch of all deep learning models.

K. More Visualizations

Task-Based Brain Region Importance To better understand the roles of each brain region in cognition under different fMRI tasks, we study the frequency at which each region appears in a hyperedge out of all identified hyperedges. The frequency, which can be considered as a measure of region importance, is visualized in Figure 10. Visual regions (*Fusiform*, *Cuneus*, *Calcarine*) are especially active due to the intensive visual demands in all three tasks. We found that the *Inferior frontal gyrus, opercular part* and the *Anterior cingulate cortex, pregenual*, recognized for their participation in response inhibition (Pornpattananangkul et al., 2016), frequently appear in the SST task. This aligns with what the SST task was designed to test. Interestingly, of the three tasks (SST, EN-back, MID), only EN-back prominently involves the *Amygdala*, a key region for emotion processing. This makes sense as EN-back is the only task related to emotion.

Resting-State Brain Region Importance We visualize the region importance of the resting state in Figure 11. Different from task states, where specific brain regions are activated in response to particular tasks, brain activities during resting states, are not driven by external tasks, leading to more diffuse and less predictable patterns of activation. This makes it harder to pinpoint specific interactions or functions.

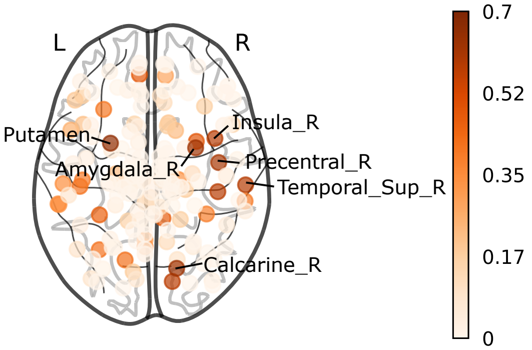


Figure 11: Region importance of resting state.

L. Discussion of the Possible Methods to Interpret Hyperedges

As mentioned in the main text, high-order relationships are much harder to interpret than pairwise ones given the exponential complexity. Here we propose a potential hierarchical strategy that tries to interpret them.

- **Adaptive Clustering Algorithm with Adjustable Granularity**: Suppose we have a clustering algorithm, where the granularity (or number of clusters) can be controlled. The clustering algorithm can cluster brain regions based on their functions (e.g. motor, visual, . . .) to different function modules (Shen et al., 2010).
- **Initial Analysis at Lower Granularity**: Set the granularity at a low level (i.e. small number of clusters). In this case, the high-order relationships are much easier to interpret (since the degree is low). However, some high-order relationships will connect to the same set of clusters, thus implying the interactions between the same set of function modules. For example, high-order relations h_1 and h_2 both connect to visual, motor, and emotion.
- **Refinement through Increased Granularity**: Find the high-order relationships that connect to the same set of function modules, and increase the granularity level, so we can tell the difference between them. For example, $h1$ may connect to the left part of the visual module, while h_2 connects to the right part.