A Biologically Interpretable Graph Convolutional Network to Link Genetic Risk Pathways and Neuroimaging Phenotypes of Disease

Sayan Ghosal[1], Giulio Pergola[2,3], Qiang Chen[2], Aaron L. Goldman[2], William Ulrich[2], Daniel R. Weinberger[2], and Archana Venkataraman[1]

[1]Department of Electrical and Computer Engineering, Johns Hopkins University, USA
[2]Lieber Institute for Brain Development, USA
[3]Department of Basic Medical Sciences, Neuroscience and Sense Organs, University of Bari Aldo Moro, Italy

Abstract

We propose a novel deep neural network for whole-genome imaging-genetics. Our genetics module uses hierarchical graph convolution and pooling operations that mimic the organization of a well-established gene ontology to embed subject-level data into a latent space. The ontology implicitly tracks the convergence of genetic risk across biological pathways, and an attention mechanism automatically identifies the salient edges in our network. We couple the imaging and genetics data using an autoencoder and predictor, which couples the latent embeddings learned for each modality. The predictor uses these embeddings for disease diagnosis, while the decoder regularizes the model. For interpretability, we implement a Bayesian feature selection strategy to extract the discriminative biomarkers of each modality. We evaluate our framework on a population study of schizophrenia that includes two functional MRI (fMRI) paradigms and gene scores derived from Single Nucleotide Polymorphism (SNP) data. Using 10-fold cross-validation, we show that our model achieves better classification performance than the baselines. In an exploratory analysis, we further show that the biomarkers identified by our model are reproducible and closely associated with deficits in schizophrenia.

1 Introduction

Neuropsychiatric disorders, such as autism and schizophrenia, have two complementary viewpoints. On one hand, they are associated with behavioral and cognitive dysfunction [28, 8], which manifest through altered brain activity [27]. On the other hand, they are highly heritable, which suggests a strong genetic underpinning [17, 41]. Most studies focus on either disentangling the neural mechanisms of the disorder via neuroimaging [20] or pinpointing genetic variations via genotyping and sequencing [41]. While these studies have shed new light on psychiatric disorders, the separation between the modalities provides an incomplete picture of the underlying pathology [27, 26].

Imaging-genetics is an emerging field that tries to merge these complementary viewpoints [21]. The imaging features are often derived from structural and functional MRI (s/fMRI), and the genetic variants are typically captured by Single Nucleotide Polymorphisms (SNPs). Data-driven imaging-genetics methods can be grouped into four categories. The first category relies on sparse multivariate regression to predict one or more imaging features based on a linear combination of SNPs [49, 32]. Though simple, interpretable, and easy to implement these models do not naturally incorporate phenotypical information, such as disease status. The second category uses representation learning

Preprint. Under review.

[37, 25] to maximally align the data distributions of the two modalities. While these models capture multivariate interdependencies between the imaging and genetics data, one again, they do not account for disease status, meaning that the identified biomarkers may not be discriminative. The third category relies on generative models [4, 18] to fuse imaging and genetics data with subject diagnosis. While this fusion extracts discriminative biomarkers, the generative nature of these methods makes it hard to compensate for missing data or to add data modalities. In fact, clinical neuroscience is moving towards multimodal imaging acquisitions where different snapshots of the brain contain potentially orthogonal information about the disease. However, with multimodal data come the problem of missing data acquisitions. Thus, integrating these multimodal and partially-complete data in single framework is essential for a comprehensive understanding of neuropsychiatric disorders.

2 Related Work: Deep Learning for Biological Data Analysis

The rise of deep learning has provided a new foundation for imaging and genetics analyses [44, 54]. In fact, the fourth category of data-driven imaging-genetics uses deep learning to link the multiple viewpoints [12, 43]. For example, the seminal work of [19] uses coupled autoencoder and predictor networks to fuse multiple imaging modalities, SNP data, and diagnosis that accommodates missing modalities. While this work provides an important stepping stone, it (and, in fact, all imaging-genetics methods) requires a drastically reduced set of genetic features to ensure model stability [7]. In terms of scale, raw dataset of 100K SNPs is reduced to 1K SNPs, often based on a Genome-Wide Association Study (GWAS) [38]. In contrast, neuropsychiatric disorders are polygenetic, meaning that they are influenced by numerous genetic variants acting across many biological pathways. The GWAS sub-selection step effectively removes information about such genetic interactions [23]. Within the genetics realm, there is a vast literature that associates SNPs and genes to different biological pathways [34, 2, 11]. The works of [22, 16] have used this information to design a sparse artificial neural network (ANN) that aggregate genetic risk according to these pathways in order to predict a phenotypic variable. While an important first step, their ANN contains just a single hidden layer, which does not account for the hierarchical and interconnected nature of the biological processes.

One way of parsing the high-dimensional and interconnected nature of genetics data is to embed these relationships into the deep model. Graph convolutional networks (GCNs) [30] provide a natural means to incorporate topological information and data geometry in the form of connections between the nodes. GCNs have gained immense popularity in protein structure prediction [14], drug discovery [45], and gene-gene interactions [53]. Beyond the architectural design, graph attention layers learn a set of "edge importance" weights, which enable us to track the information flow through the graph [48]. The attention acts as an implicit regularizer by encouraging the model to focus on the most discriminative set of interactions between the nodes, often leading to better generalization on unseen data. Graph attention has been successful in biological applications, like predicting disease-RNA association [55] and essential gene prediction [42], as well as in non-biological applications, like machine translation [3] and image classification [35]. Finally, similar to max-pooling operations in standard convolutional networks, graph pooling allows us to aggregate information at each level of the hierarchy [9, 31, 52]. During standard graph pooling, the nodes at each level are clustered to form a smaller subgraph. These cluster assignments can be obtained via deterministic algorithms [9] or be learned in an end-to-end fashion [52, 31]. One drawback to this approach is that the clustering operation can obscure the interpretations of each node, as the information has now been fused. Our approach bypasses this issue by using a sparse hierarchical graph for convolution and pooling.

One drawback of deep learning is the black-box nature of the trained models. Recently, the work of [33] has provided a framework called SHAP (SHapley Additive exPlanations) to compute feature importance as a step towards interpretability. SHAP assigns a score to each feature equal to the marginal contribution of that feature to the output, as averaged over all possible feature subsets. Variants like kernel-SHAP provide for increased flexibility of the feature importance computation [33, 40]. While SHAP provides a model-agnostic tool to quantify input-output relationships, it says little about how groups of input features affect the output, which is a primary interest in biological datasets.

In this paper, we introduce an interpretable **G**enetics and m**U**ltimodal Imaging based **DE**ep neural network (**GUIDE**), for whole-brain and whole-genome analysis. We use hierarchical graph convolutions that mimic an ontology based network of biological processes [34]. Using pooling operation we capture the information flow through the graph. Finally, we use graph attention to learn the salient edges of the graph, thus identifying key *pathways* associated with the disorder. On the imaging front,

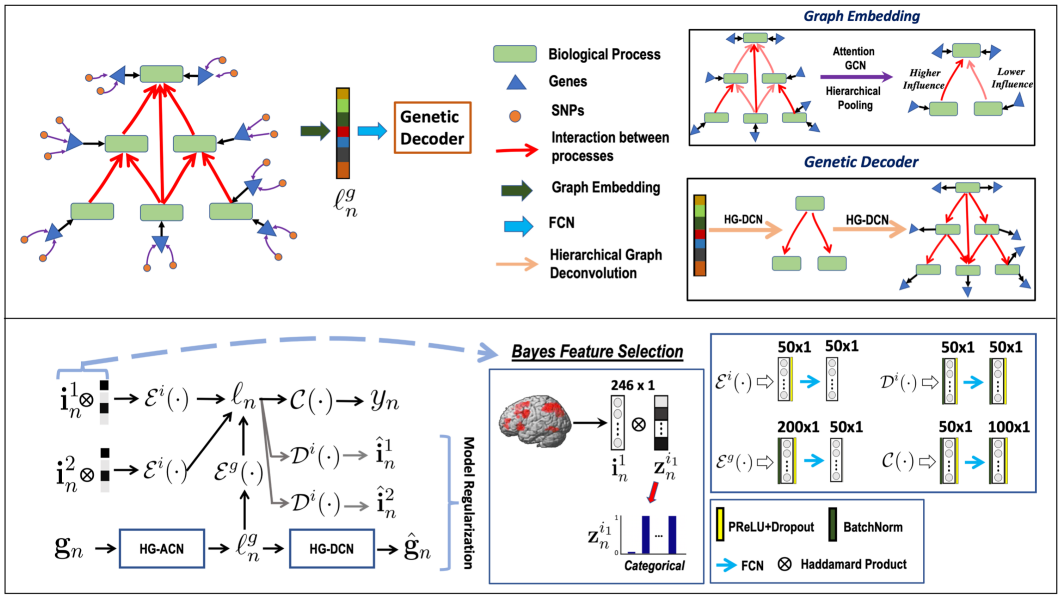

Figure 1: Overview of the GUIDE framework. **Top:** Gene embedding using attention based hierarchical graph convolution. We also depict the unpooling operation used as a regularizer. **Bottom:** Imaging and genetics integration; both modalities are coupled for disease classification. The variables $\{\mathbf{i}_n^1, \mathbf{i}_n^2\}$ correspond to the imaging data, and $\mathbf{g}_n$ is the genetic data. $\mathcal{E}(\cdot)$, $\mathcal{D}(\cdot)$, $\mathcal{C}(\cdot)$ are the feature extraction, model regularization, and classification operations, respectively.

we use an autoencoder, coupled with Bayesian feature selection to learn multivariate importance scores. The latent embeddings learned by our genetics graph convolutions and imaging autoencoder are combined for disease prediction. We evaluate GUIDE on a population study of schizophrenia that includes two task fMRI paradigms and SNP data. Our framework achieves better classification performance than the baselines and decoupled (imaging or genetics only) variants of our model. We also demonstrate that it identifies reproducible and clinically relevant phenotypes of schizophrenia.

3 A Deep Multimodal Imaging-Genetics Framework

Fig. 1 illustrates our GUIDE framework. Inputs are the gene scores $\mathbf{g}_n \in \mathbf{R}^{G\times 1}$ for each subject n and the corresponding imaging feature vectors $\mathbf{i}_n^1 \in \mathbf{R}^{M_1\times 1}$ and $\mathbf{i}_n^2 \in \mathbf{R}^{M_2\times 1}$ obtained from two different acquisitions. The gene scores $\mathbf{g}_n$ are obtained by grouping the original SNPs according to the nearest gene and aggregating the associated genetic risk weighted by GWAS effect size [38]. During training, we know the subject diagnosis $y_n \in \{0, 1\}$, but this data is absent during testing.

3.1 Genetic Graph Embedding

The top portion of Fig. 1 illustrates our attention-based graph convolution model for genetic embedding. The underlying graph is based on the gene ontological hierarchy [2] that effectively maps each gene to different biological processes. Mathematically, this ontology gives rise to a sparse matrix $\mathbf{A}_g \in \{0, 1\}^{G\times T}$, where G is the total number of genes, and T is the number of biological processes (i.e., nodes) encoded by the ontology. Given this substrate, GUIDE first learns a projection of the gene scores $\mathbf{g}_n$ onto a d-dimensional signal $\mathbf{p}_n(t) \in \mathbf{R}^{1\times d}$ at each node $t \in \{1, \ldots, T\}$:

$$\mathbf{p}_n(t) = PReLU(\mathbf{g}_n^T \mathbf{W}_t \otimes \mathbf{A}_g[:, t]), \tag{1}$$

where each of the columns of the learned weight matrix $\mathbf{W}_t \in \mathbf{R}^{G\times d}$ are masked by the nonzero entries in the tth column of $\mathbf{A}_g$. We note that that this projection step is similar to a single-layer perceptron with only a subset of input nodes connected to each hidden node.

3.2 Graph Attention and Hierarchical Pooling

In addition to grouping genes into biological processes, a standard ontology also specifies a hierarchical relationship between the biological processes themselves [1]. An example of this hierarchy is: generation of neurons $\rightarrow$ neurogenesis $\rightarrow$ nervous system development.

We use a graph convolution to mimic the flow of information through the hierarchy of biological processes. Here, GUIDE learns two complementary pieces of information. The first is a set of graph convolutional filters that act on each embedded node signal, and the second is a set of subject-specific attention weights that select the discriminative edges. Formally, let the binary matrix $\mathbf{A}_p \in \mathbf{R}^{T \times T}$ capture the directed edges in the ontology. Our graph convolution at stage l is:

$$\mathbf{p}_n^{l+1}(t) = \sigma\left(\sum_{j \in Child(t)} \mathbf{E}_n^l(t,j)\mathbf{p}_n^l(j)\mathbf{W}^l + \beta_t \mathbf{p}_n^l(t)\mathbf{W}_s\right), \tag{2}$$

where $\mathbf{p}_n^l(t) \in \mathbf{R}^{1 \times d_l}$ is signal for node t at stage l, $\mathbf{W}^l \in \mathbf{R}^{d_l \times d_{l+1}}$ is the convolutional filter between stages l and $l+1$, β_t is the self-influence for node t, $\mathbf{W}_s \in \mathbf{R}^{d_l \times d_{l+1}}$ is the convolution filter for self loop, and $\sigma(\cdot)$ is the nonlinearity. The summation in Eq. (2) aggregates the influence over all child nodes defined by the graph $\mathbf{A}_p$, thus respecting the high-level ontology.

The variables $\mathbf{E}_n^l(t,j)$ quantifies the influence of the child node j over the parent node t at convolutional stage l. Unlike a standard graph convolutional network, in which the edge weights are fixed, we learn $\mathbf{E}_n^l(t,j)$ using a graph attention mechanism. Mathematically, we have

$$\mathbf{E}_n^l(t,j) = \frac{\exp\left(\tanh\left(\left[\mathbf{p}_n^l(t)\mathbf{W}^l \quad \mathbf{p}_n^l(j)\mathbf{W}^l\right] \cdot \mathbf{c}^l\right)\right)}{\sum_{j \in Child(t)} \exp\left(\tanh\left(\left[\mathbf{p}_n^l(t)\mathbf{W}^l \quad \mathbf{p}_n^l(j)\mathbf{W}^l\right] \cdot \mathbf{c}^l\right)\right)} \tag{3}$$

where $\mathbf{c}^l$ is a fixed weight vector learned during training. We estimate the self influence variable as $\beta_t = \sigma\left(\mathbf{p}_n^l(t)\mathbf{W}_s \cdot \mathbf{c}_s^l\right)$, where $\mathbf{c}_s^l$ is another weight vector learned during training.

Finally, we use hierarchical pooling to coarsen the graph. As formulated, the graph convolution passes information "upwards" from child nodes to parent nodes. From the ontology standpoint, each stage of the hierarchy goes from a lower-level biological process (e.g., neurogenesis) to a higher-level biological process (e.g., nervous system development). Thus, we remove the lowest (leaf) layer from the graph at each stage l and continue this process until we reach the root nodes. The final encoding $\ell_n^g \in \mathbf{R}^{\mathcal{D} \times 1}$ is obtained by concatenating the signals from each root node. We refer to our genetics module as a Hierarchical Graph Attention Convolution Network (HG-ACN).

3.3 Bayesian Feature Selection

The bottom branch of Fig. 1 shows our encoding and feature selection procedure for the imaging modalities $\mathbf{i}_n^1$ and $\mathbf{i}_n^2$ of each subject n. From a Bayesian viewpoint, the problem of feature selection can be handled by introducing a latent binary random vector $\mathbf{z}_n^m$ of the same dimensionality as $\mathbf{i}_n^m$ and inferring its posterior probability distribution given the paired training dataset: $\mathcal{D} = \{\mathbf{i}_n^m, y_n\}$, where y_n denotes the class label. By defining $\mathbf{I}^m = [\mathbf{i}_1^m, \ldots, \mathbf{i}_n^m]$, $\mathbf{Y} = [y_1, \ldots, y_n]$, and $\mathbf{Z}^m = [\mathbf{z}_1^m, \ldots, \mathbf{z}_n^m]$, we note that the desired posterior distribution $p(\mathbf{Z}^m|\mathbf{I}^m, \mathbf{Y})$ is intractable. One strategy is to minimize the KL divergence between an approximate distribution $q(\cdot)$ and the true posterior distribution $KL\left(q(\mathbf{Z}^m)||p(\mathbf{Z}^m|\mathbf{I}^m, \mathbf{Y})\right)$. Mathematically, this is equivalent to

$$\underset{q(\cdot)}{\operatorname{argmin}} \quad -E_q\left[\log\left(p\left(\mathbf{Y}|\mathbf{I}^m, \mathbf{Z}^m\right)\right)\right] + KL\left(q(\mathbf{Z}^m)||p(\mathbf{Z}^m)\right), \tag{4}$$

where $p(\mathbf{Z}^m)$ is a prior distribution over the latent masks. While Eq. (4) does not have a closed-form solution, it can be optimized via Monte Carlo integration by sampling the vectors $\mathbf{z}_n^m \sim Bernoulli(\mathbf{b}^m)$, where $\mathbf{b}^m$ parameterizes the approximate distribution $q(\cdot)$, and minimizing the empirical form of Eq. (4) [15, 29]. In this case, the first term becomes the binary cross entropy loss where the input features $\mathbf{i}_n^m$ are masked according to $\mathbf{z}_n^m$. In order to learn the probability maps $\mathbf{b}^m$ during training, we replace the binary $\mathbf{z}_n^m$ with a continuous relaxation of the Bernoulli distribution:

$$\mathbf{z}_n^m = \sigma\left(\frac{\log(\mathbf{b}^m) - \log(1-\mathbf{b}^m) + \log(\mathbf{u}_n^m) - \log(1-\mathbf{u}_n^m)}{r}\right), \tag{5}$$

where $\mathbf{u}_{i_1}^n$ is a random vector sampled from $Uniform(0,1)$, the parameter r controls the relaxation from the Bernoulli distribution and $\mathbf{b}^m$ captures the feature selection probabilities for modality m.

During training, we sample the random variables $\mathbf{z}_n^1, \mathbf{z}_n^2$ during each forward pass and use it to mask the input data for patient n. We implicitly optimize the empirical form of Eq. (4) via backpropagation on the standard cross-entropy loss of the predicted versus true class label.

3.4 Multimodal Fusion and Model Regularization

As shown in Fig. 1, the imaging data $\mathbf{i}_n^m$ and feature mask $\mathbf{z}_n^m$ are passed through an encoder $\mathcal{E}^m(\cdot)$ and projected to a low-dimensional latent space. We implement this encoder using a standard ANN. We fuse the latent embeddings across modalities to obtain a common representation:

$$\ell_n = \frac{1}{M_n}\left(\mathcal{E}^1(\mathbf{i}_1^n, \mathbf{z}_n^1) + \mathcal{E}^2(\mathbf{i}_n^2, \mathbf{z}_n^2) + \text{HG-ACN}(\mathbf{g}_n)\right), \tag{6}$$

where M_n is the number of modalities present for subject n. This fusion operation enables our model to handle missing data and leverages the correlation structure shared between the modalities to identify predictive biomarkers. We pass this low dimensional latent representation through a classifier and minimize a binary cross entropy loss for diagnostic prediction of y_n.

We introduce three regularizers to stabilize the model. The genetic regularizer reconstructs the gene scores by performing a hierarchical graph deconvolution (HG-DCN) as shown in Fig. 1. This operation unwraps the gene encoding via the same ontology hierarchicy. The imaging data is reconstructed though another ANN denoted as $\mathcal{D}(\cdot)$. Our joint objective function can be written:

$$\begin{aligned}\mathcal{L}(\mathbf{i}_1, \mathbf{i}_2, \mathbf{g}) = &-\sum_{n=1}^{N}\left(y_n \log(\hat{y}_n) + (1-y_n)\log(1-\hat{y}_n)\right)\\ &+\sum_{m=1}^{2} KL_{\mathcal{M}}\left(Ber(\mathbf{b}^m)||Ber(\mathbf{s})\right) + \lambda_I \sum_{n=1}^{N_1}||\mathbf{i}_n^1 - \mathcal{D}^1(\ell_n)||_2^2 + \lambda_I \sum_{n=1}^{N_2}||\mathbf{i}_n^2 - \mathcal{D}^2(\ell_n)||_2^2\\ &+\lambda_G \sum_{n=1}^{N_g}||\mathbf{g}_n - \text{HG-DCN}(\ell_n^g)||_2^2\end{aligned} \tag{7}$$

where N is the total number of subjects, and the hyper-parameters $\{\lambda_I, \lambda_G\}$ control the contributions of the data reconstruction errors. The prior over $\mathbf{z}_n^m$ appears as the KL divergence between the learned distribution $Ber(\mathbf{b}^m)$ and a binary random vector $Ber(\mathbf{s})$ with small entries to enforce sparsity. The function $KL_{\mathcal{M}}(\cdot||\cdot)$ in Eq. 7 averages the element-wise KL divergences across the input feature dimension, thus maintaining the scale of the prior term regardless of dimensionality. If a data modality is missing for subject n, we simply fix that encoder-decoder branch in Fig. 1 (equivalent to removing the reconstruction loss in Eq. (6)) and train the rest of the model using backpropagation.

Implementation Details: Our main cohort consists of 1848 subjects with only genetics data and an additional 208 subjects with both imaging and genetics data. Accordingly, we pretrain the HG-ACN and HG-DCN branches and classifier of GUIDE using the 1848 genetics-only subjects. These 1848 subjects are divided into a training and validation set, the latter of which is used for early stopping criteria. We use the pretrained model to warm start our GUIDE framework and perform 10-fold nested cross validation over the 208 imaging-genetics cohort. We perform a grid search over three order of magnitude and fix the hyperparameters for all experiments at: $\lambda_I = 3 \times 10^{-3}, \lambda_G = 10^{-5}$. We use a signal dimensionality of $d = 2$ for the gene embedding and $d_l = 5$ for the subsequent graph convolution and deconvolution operations. The non-linearity in Eq. (2) is given by a $LayerNorm$, followed by $PReLU$ and $Dropout$. The Bernoulli prior over $\mathbf{z}^m$ is set at $s = 0.001$. The remaining architectural details are shown in Fig. 1. We use ADAM with an initial learning rate of 0.0002 and decay of 0.7 every 50 epochs. Our code is implemented using Matlab 2019b and PyTorch 3.7. Training the full model takes roughly 17hrs on a 4.9GB Nvidia K80 GPU.

3.5 Baseline Comparison Methods

Parallel ICA + RF: We concatenate the imaging modalities to single vector $\mathbf{i}_n = [\mathbf{i}_n^{1^T} \quad \mathbf{i}_n^{2^T}]^T$ and perform parallel ICA (p-ICA) [37] with the gene scores $\mathbf{g}_n$. Since p-ICA cannot handle missing

modalities, we fit a multivariate regression model to impute a missing imaging modality from the available ones. Specifically, if $\mathbf{i}_1^n$ is absent, we impute it as: $\mathbf{i}_1^n = \boldsymbol{\beta}\,\mathbf{i}_2^n$, where $\boldsymbol{\beta}$ is the regression coefficient matrix obtained from training data. After imputation, we use p-ICA to decompose the imaging and genetics data into independent but interrelated networks:

$$\mathbf{i}_n = \mathbf{S}\,\mathbf{e}_n \quad \text{and} \quad \mathbf{g_n} = \mathbf{W}\,\mathbf{f}_n$$

where $\mathbf{S}, \mathbf{W}$ are independent source matrices and the $\mathbf{e}_n, \mathbf{f}_n$ are loading vectors. We concatenate the loading matrices $\left[\mathbf{e}_n^T, \mathbf{f}_n^T\right]$ and use it as the input feature vector for a random forest classifier.

During training, we apply p-ICA to just the training data to estimate the sources $\{\mathbf{S}_{train}, \mathbf{W}_{train}\}$. We use these estimated sources to obtain the loading matrices for the test data as follows:

$$\mathbf{i}_{test} = \mathbf{S}_{train}\mathbf{e} \quad \text{and} \quad \mathbf{g}_{test} = \mathbf{W}_{train}\mathbf{f}$$

Our hyperparameter tuning optimizes the number and depth of the decision trees. We controlled the depth of the tree depth by setting the minimum number of observations per leaf node. We perform a grid search to fix these parameters at $\{\text{No. trees} = 10000, MinleafSize = 50\}$.

G-MIND: The model from [19] was designed for a low-dimensional (1242) dimensional genetics input vector. Since this work focuses on whole-genome analysis, we introduce a linear fully-connected layer to project the high-dimensional gene scores $\mathbf{g}_n$ to a 1242 dimensional vector for input to G-MIND. The hyperparameters are fixed according to the approach in [19]. We evaluate both random weight initialization and pretraining the genetics branch with the same 1848 subjects used for GUIDE.

Ablation Models: We consider two ablated versions of GUIDE. The first consists of the genetics branches and classifier, and the second consists of just the imaging branches and classifier. We optimize these networks architectures using nested 10-fold cross validation outlined in Section 3.4.

3.6 Evaluation Strategy

We conduct a comprehensive evaluation of our framework that includes classification performance (i.e., patient versus control prediction) and biomarker reproducibility.

Classification Performance: We use repeated 10-fold cross validation on the 208 imaging-genetics subjects to quantify the classification performance. The models are trained using 8 folds, and the remaining two folds are reserved for validation and testing. We report accuracy, sensitivity, specificity, Area Under the ROC Curve (AUROC), and Area Under the precision-recall Curve (AUPRC). The operating point is chosen by minimizing the classification error on the validation set.

Influence of the Ontology-Based Hierarchy: One implicit assumption of GUIDE is that processing whole-genome data according to an established gene ontology will allow us to strategically consolidate the high-dimensional information for biomarker discovery and class prediction. To test this assumption, we compare our gene ontology network to random networks and an unstructured model. The random network contains the same number of nodes in each layer as the ontology-based graph. However, we randomly permute edges between the nodes from one layer to the next and between the genetic inputs $\mathbf{g}_n$ and the network nodes. The corresponding unstructured model is a fully-connected ANN with the same number of parameters but no inherent structure between layers.

We train the ontology-based network, the randomized networks and the ANN using the 1848 genetics-only subjects and test them over the remaining 208 subjects. In our experiment we also sample 10 random graph and evaluate their performances. This experiment quantifies the value of preserving the biological hierarchy over a similar architecture without this prior knowledge.

Reproducibility of Feature Importance Maps: The feature probability maps $\mathbf{b}_m$ capture the importance of the features of modality m for the downstream task. Using the repeated 10-fold cross validation setup, we extract the top-K features learned by the importance maps and represent this information in the binary encoding vectors $\mathbf{r}_f^m$, where $\mathbf{r}_f^m(i) = 1$ if i is one of the top K regions for fold f and modality m. We then mask our test data from each fold send it through GUIDE for patient versus control classification. We compare this masking procedure with Kernel SHAP [33]. The background values for K-SHAP are fixed to the average input data across the training set. We

Table 1: Classification performances (mean $\pm$ std) across ten 10-fold cross validation runs.

Method \ Perf	Sensitivity	Specificity	Accuracy	AUROC	AUPRC
P-ICA	0.30 ± 0.18	$\underline{0.74 \pm 0.18}$	0.54 ± 0.03	0.55 ± 0.03	0.51 ± 0.03
GMIND (random)	0.45 ± 0.11	0.68 ± 0.11	0.60 ± 0.03	0.60 ± 0.03	0.54 ± 0.04
GMIND (pretrain)	0.51 ± 0.13	0.63 ± 0.11	0.57 ± 0.03	0.61 ± 0.02	0.55 ± 0.02
Imaging Only	$\underline{0.52 \pm 0.18}$	0.62 ± 0.15	0.57 ± 0.03	0.60 ± 0.03	0.53 ± 0.02
Genetic Only	0.47 ± 0.16	$\mathbf{0.76 \pm 0.11}$	$\underline{0.63 \pm 0.02}$	$\underline{0.68 \pm 0.02}$	$\mathbf{0.63 \pm 0.02}$
GUIDE	$\mathbf{0.57 \pm 0.07}$	0.71 ± 0.05	$\mathbf{0.64 \pm 0.02}$	$\mathbf{0.69 \pm 0.02}$	$\underline{0.61 \pm 0.02}$

use the validation set to select the top K K-SHAP features and use them to mask the testing data for classification through GUIDE . The binary masks learned through K-SHAP are denoted as $\mathbf{k}_f^m$, where $\mathbf{k}_f^m(i) = 1$ if i is one of the top K regions for fold f and modality m.

To quantify the reproducibility of the top-K features, we calculate the pairwise $cosine$ similarity between all the binary vectors across the folds as identified by our Bayesian approach and K-SHAP. The distribution of similarities tells us how often the same regions appear across subsets of the data.

Discovering of Biological Pathways: The attention layer of HG-ACN provides a subject-specific measure of "information flow" through the network. To identify discriminative pathways, we first trace all possible paths between leaf and root nodes. The importance of edges along these paths are used in a logistic regression framework to predict the subject class label. We then perform a likelihood ratio test and obtain a p-value (FDR corrected) for each path in terms of patient/control differentiation. For robustness, we repeat this experiment 10 times using subsets of 90% of our total dataset $1848 + 208$ subjects and select pathways that achieve $p < 0.05$ in at least 7 of the 10 subsets.

4 Results

Data and Preprocessing: Genotyping was done using variate Illumina Bead Chips including 510K/610K/660K/2.5M. Quality control and imputation were performed using PLINK and IMPUTE2 respectively. The resulting 102K linkage disequilibrium independent SNPs are grouped to nearest gene (within 50kb basepairs) [50]; the 13908 dimensional input $\mathbf{g}_n$ is computed by averaging the genetic risk from the GWAS study of [38] across the corresponding SNPs. We use the first five layers of the gene ontology network [2] to construct our directed graph. In total, this network encompasses 13595 biological processes organized from 2836 leaf nodes to 3276 root nodes.

Our imaging dataset includes two task fMRI paradigms from 208 subjects provided by Anonymous Institution (ANO). The first fMRI paradigm is a working memory task (Nback). During the 0-back blocks, participants were instructed to press a button corresponding to a number displayed on the screen. During the 2-back working memory blocks, participants were instructed to recall the number they had seen two stimuli previously. The second paradigm is a simple declarative memory task (SDMT) which involves incidental encoding of complex aversive visual scenes. FMRI data was acquired on 3T GE Sigma scanner (EPI, TR/TE = 2000/28 msec; res $= 3.75 \times 3.75 \times 6/5$ for NBack and SDMT, respectively). FMRI preprocessing includes slice timing correction, realignment, spatial normalization to an MNI template, smoothing and motion regression. We use SPM 12 to generate contrast maps for each paradigm and the Brainnetome atlas [13] to define 246 cortical and subcortical regions. The inputs $\mathbf{i}_n^1, \mathbf{i}_n^2$ correspond to the average contrast over voxels in the respective region.

In total, our dataset contains 1848 subjects with just SNP data, 98 subjects with SNP and N-back data, 48 subjects with SNP and SDMT data, and 62 subjects who have all the three data modalities.

Model Performance: Table 1 reports the 10-fold cross validation (CV) testing performance of all the methods on the multimodal imaging-genetics dataset. We repeat CV ten times with random distribution of the dataset to quantify the distribution of the performance metrics. We note that P-ICA and GMIND with random initialization have relatively poor performance, likely due to the high

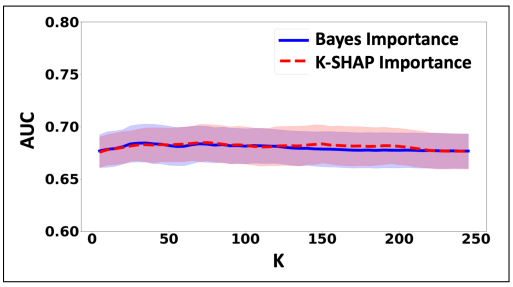

Figure 2: Mean AUC and confidence interval when masking the top-K imaging features learned by our Bayesian feature selection procedure (solid blue) and K-SHAP (dashed red). K is varied along the x-axis.

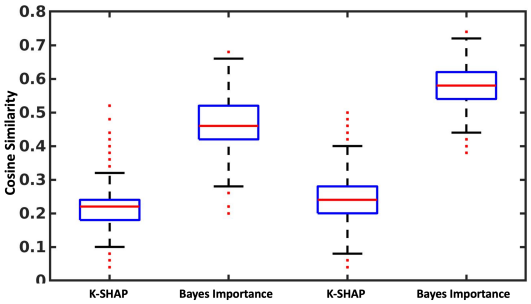

Figure 3: Distribution of pairwise *cosine* similarities between the feature masks r_m^f (Bayesian feature selection) and k_m^f (K-SHAP) across the repeated CV folds.

dimensionality of $\mathbf{g}_n$ and low sample size. Pretraining on a separate genetics dataset improves the AUC of GMIND, highlighting the benefits of increased data. The Imaging Only performance is also poor, possibly due to the missing fMRI acquisitions. GUIDE achieves the best performance, which can be attributed to the hierarchical information flow through the ontology graph and multimodal fusion. Additionally, our Genetics Only model performs similar to GUIDE, which demonstrates that the biologically inspired neural network can extract predictive features from whole-genome data.

Benefit of the Ontology over Random Graphs and Unstructured Networks: Fig. 4 illustrates the classification ROC computed over both training (left) and testing (right) data. As seen, the training performance is similar for all the three models. However, the testing performance is near-chance using both random graphs and the ANN. In contrast, our gene ontology network (blue line in Fig. 4) is able to generalize to new data, which suggests that our biologically-inspired ontology network can extract more robust and meaningful genetic biomarkers of the disorder.

Feature Importance Maps: Fig. 2 illustrates the classification AUC when the input features are masked according to the top-K importance scores learned by our Bayesian feature selection maps $\{\mathbf{b}_f^1, \mathbf{b}_f^2\}$ (solid blue) and by K-SHAP (dashed red). In both cases, we have pretrained the genetic branch and the classifier of GUIDE on 1848 genetics-only dataset and use it to warm start our repeated 10-fold cross validation experiments on the 208 imaging-genetics data.

As seen, both feature selection schemes achieve similar AUCs as the number of features K is varied across its entire range, thus highlighting the robustness of our (simpler) approach.

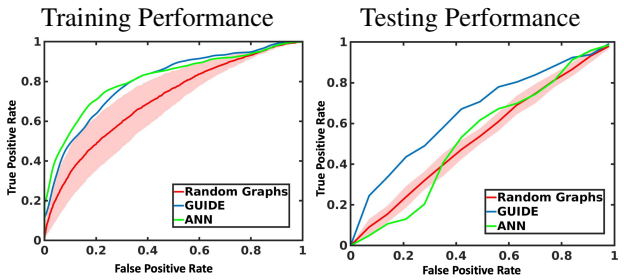

Figure 4: ROCs for unstructured ANN (green) and the structured models where the HG-ACN and HG-DCN use either random graphs (red) or the gene ontology network (blue).

Fig. 3 illustrates the distribution of *cosine* similarities between the masked feature vectors learned by each method (Bayesian feature selection and K-SHAP) across the multiple 10-fold cross validation training folds. Here, we have fixed $K = 50$, and we run the cross validation procedure 10 times, which yields 100 total folds and 4950 pairwise comparisons per method. We have also separated the N-back and SDMT modalities in Fig. 3. Notice that our Bayesian feature selection procedure achieves significantly higher *cosine* similarity values, which suggests that it selects a more robust set of features that is consistent across subsets of our main cohort. Whereas K-SHAP draws inference on individual level data which varies largely across random subsets of the population.

Fig. 5 illustrates the imaging features that are most consistently selected by our method across the folds for $K = 50$. Recall that each feature corresponds to a brain region, which we have colored according to the selection frequency. For clarity, we have displayed only the top 30% regions. We

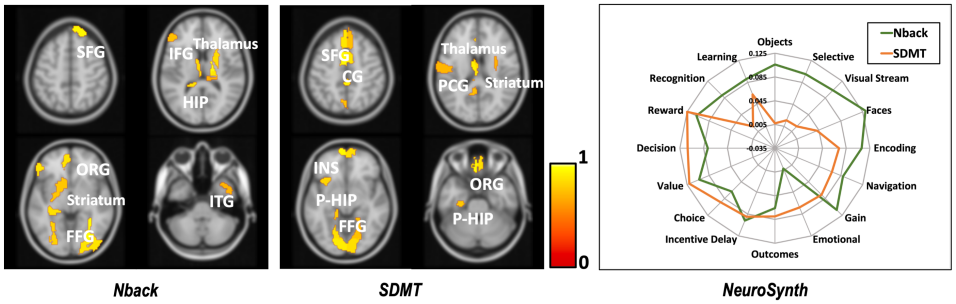

Figure 5: **Left:** The consistent set of brain regions captured by the dropout probabilities $\{\mathbf{b}^1, \mathbf{b}^2\}$ for $K = 50$. The color bar denotes the selection frequency. **Right:** Brain states associated with the selected regions for each fMRI task, as decoded by Neurosynth.

observe that the N-back biomarkers involve the Superior Frontal Gyrus (SFG) and Inferior Frontal Gyrus (IFG), which are associated with executive cognition [10]. The SDMT biomarkers implicate the parahippocampal gyrus, SFG, and default mode network. For additional interpretation, we decode the higher order brain states implicated by these regions using Neurosynth [51]. Neurosynth uses a reverse-inference procedure to select a set of "cognitive terms" associated with a set of input coordinates based on how frequently similar patterns have been observed across the fMRI literature. Fig. 5 shows that the N-back and SDMT biomarkers are associated with memory encoding and decision making [39, 24], thus verifying that GUIDE captures information relevant to the fMRI tasks.

Genetic Pathways: GUIDE contains 14881 "biological pathways" between the leaf and root nodes. Using the likelihood ratio testing procedure outlined in Section 3.6, we are able to identify 152 pathways with $p < 0.05$ after FDR correction that appear in at least 7 of the 10 random training datasets. We cluster these pathways into 10 categories based on their semantic similarity. Here, we use the tf-idf [5] information retrieval scheme to extract keywords in the pathways; we then embed them in a two-dimensional space using t-SNE [46] and apply a simple k-means clustering algorithm. Fig. 6 shows the clusters along with the most frequent keywords within each cluster. As seen, the frequent biological processes involve calcium signaling, regulation of macrophage and immunological synapse formation which have been previously linked to schizophrenia [36, 6, 47]. We note that this is an exploratory analysis that requires further verification. However, this experiment shows that GUIDE can be used to extract discriminative biological information about neuropsychiatric disorders.

5 Conclusion

We propose a novel biologically interpretable graph convolutional network that integrates imaging and genetics data to perform disease prediction. This model is able to leverage prior biological information of different connected biological processes to identify patterns from both imaging and genetic data

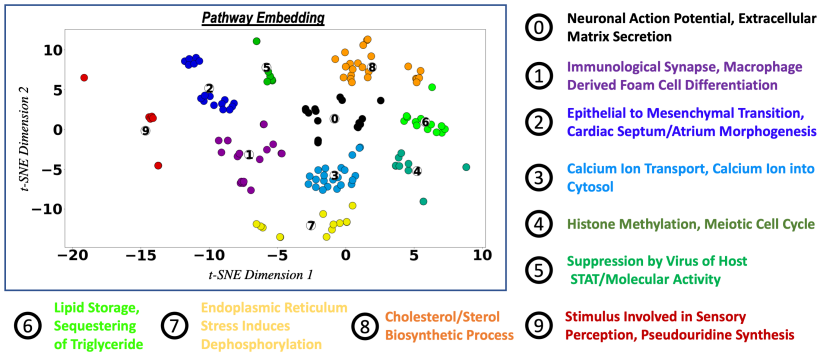

Figure 6: Ten different categories of pathways based on their semantic similarity. The key words show the most frequent biological processes within each cluster.

sets. Additionally, the unique use of Bayesian feature selection is able to find a set of clinically relevant biomarkers. The improved classification performance shows the ability of this model to build a comprehensive view about the disorder based on the incomplete information obtained from different modalities. We note that, our framework can easily be applied to other imaging modalities, such as structural and diffusion MRI, simply by adding encoder-decoder branches. In future, we will apply our framework to other application domains, such as autism and Parkinson's disease.

Acknowledgments and Disclosure of Funding

This work was supported by NSF CRCNS 1822575, and the National Institute of Mental Health extramural research program.

References

[1] Alcocer-Cuarón, C., et al.: Hierarchical structure of biological systems: A bioengineering approach. Bioengineered **5**(2), 73 (2014)

[2] Ashburner, M., et al.: Gene ontology: Tool for the unification of biology. Nature Genetics **25**(1), 25–29 (2000)

[3] Bahdanau, D., Cho, K.H., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. International Conference on Learning Representations, ICLR (2015)

[4] Batmanghelich, N.K., et al.: Probabilistic Modeling of Imaging, Genetics and Diagnosis. IEEE transactions on medical imaging **35**(7), 1765–79 (2016)

[5] Beel, J., et al.: Research-paper recommender systems: a literature survey. International Journal on Digital Libraries **17** (2016)

[6] Berridge, M.J.: Dysregulation of neural calcium signaling in Alzheimer disease, bipolar disorder and schizophrenia. In: Prion. vol. 7. Landes Bioscience (2013)

[7] Bhattacharjee, S., et al.: A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. American Journal of Human Genetics **90**(5), 821–835 (2012)

[8] Bowie, C.R., et al.: Cognitive deficits and functional outcome in schizophrenia. Neuropsychiatric Disease and Treatment **2**(4), 531–536 (2006)

[9] Bruna, J., et al.: Spectral Networks and Locally Connected Networks on Graphs. 2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings (dec 2013)

[10] Callicott, J.H., et al.: Abnormal fMRI response of the dorsolateral prefrontal cortex in cognitively intact siblings of patients with schizophrenia. American Journal of Psychiatry **160**(4), 709–719 (2003)

[11] Carbon, S., et al.: The Gene Ontology resource: Enriching a GOld mine. Nucleic Acids Research **49**(D1), D325–D334 (2021)

[12] Eraslan, G., et al.: Deep learning: new computational modelling techniques for genomics. Nature Reviews Genetics **20** (jul 2019), `https://pubmed.ncbi.nlm.nih.gov/30971806/`

[13] Fan, L., et al.: The Human Brainnetome Atlas: A New Brain Atlas Based on Connectional Architecture. Cerebral Cortex **26**(8), 3508–3526 (2016)

[14] Fout, A., et al.: Protein Interface Prediction using Graph Convolutional Networks. Advances in Neural Information Processing Systems **30** (2017)

[15] Gal, Y., Kendall, A.: Concrete Dropout

[16] Gaudelet, T., et al.: Unveiling new disease, pathway, and gene associations via multi-scale neural network. PLOS ONE **15**(4), e0231059 (2020)

[17] Gejman, P.V., et al.: The role of genetics in the etiology of schizophrenia. Psychiatric Clinics of North America **33**(1), 35–66 (2010)

[18] Ghosal, S., et al.: Bridging Imaging, Genetics, and Diagnosis in a Coupled Low-dimensional Framework. In: MICCAI: Medical Image Computing and Computer Assisted Intervention. vol. 11767 LNCS, pp. 647–655. Springer (2019)

[19] Ghosal, S., et al.: G-MIND: an end-to-end multimodal imaging-genetics framework for biomarker identification and disease classification. In: Medical Imaging 2021: Image Processing. vol. 11596, p. 8. SPIE (2021)

[20] Gur, R.E., et al.: Functional magnetic resonance imaging in schizophrenia. Dialogues in Clinical Neuroscience **12**(3), 333–343 (2010)

[21] Hariri, A., Weinberger, D.: Imaging genomics. British Medical Bulletin **65**(1), 259–270 (2003)

[22] van Hilten, A., et al.: GenNet framework: Interpretable neural networks for phenotype prediction. bioRxiv p. 2020.06.19.159152 (2020)

[23] Horváth, S., et al.: Schizophrenia as a disorder of molecular pathways. Biological Psychiatry **77**(1), 22–28 (2015)

[24] Hostiuc, S., et al.: Testing decision-making competency of schizophrenia participants in clinical trials. A meta-analysis and meta-regression. BMC psychiatry **18** (jan 2018)

[25] Hu, W., et al.: Adaptive Sparse Multiple Canonical Correlation Analysis With Application to Imaging (Epi)Genomics Study of Schizophrenia. IEEE Transactions on Biomedical Engineering **65**(2), 390–399 (2018)

[26] Karam, C.S., et al.: Signaling pathways in schizophrenia: Emerging targets and therapeutic strategies. Trends in Pharmacological Sciences **31**(8), 381–390 (2010)

[27] Karlsgodt, K.H., et al.: Structural and functional brain abnormalities in schizophrenia. Current Directions in Psychological Science **19**(4), 226–231 (2010)

[28] Kay, S.R., et al.: The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophrenia Bulletin **13**(2), 261–276 (1987)

[29] Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes

[30] Kipf, T.N., Welling, M.: Semi-Supervised Classification with Graph Convolutional Networks. 5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings (sep 2016)

[31] Lee, J., et al.: Self-Attention Graph Pooling. 36th International Conference on Machine Learning, ICML 2019 (apr 2019)

[32] Liu, J., et al.: A Review of Multivariate Analyses in Imaging Genetics. Frontiers in Neuroinformatics **8**, 29 (2014)

[33] Lundberg, S.M., et al.: A Unified Approach to Interpreting Model Predictions. In: Advances in Neural Information Processing Systems. vol. 30 (2017)

[34] Mi, H., et al.: Large-scale gene function analysis with the panther classification system. Nature Protocols **8**(8), 1551–1566 (2013)

[35] Mnih, V., et al.: Recurrent Models of Visual Attention. Advances in Neural Information Processing Systems **3** (2014)

[36] Ormel, P.R., et al.: Characterization of macrophages from schizophrenia patients. npj Schizophrenia **3**(1), 41 (2017)

[37] Pearlson, G.D., et al.: An introductory review of parallel independent component analysis (p-ICA) and a guide to applying p-ICA to genetic data and imaging phenotypes to identify disease-associated biological pathways and systems in common complex disorders. Frontiers in Genetics **6** (2015)

[38] Purcell, S.M., et al.: Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature **460**(7256) (2009)

[39] Rasetti, R., et al.: Altered hippocampal-parahippocampal function during stimulus encoding:A potential indicator of genetic liability for schizophrenia. JAMA Psychiatry **71** (2014)

[40] Ribeiro, M.T., et al.: "Why should i trust you?" Explaining the predictions of any classifier. pp. 1135–1144. Association for Computing Machinery (aug 2016), `http://dx.doi.org/10.1145/2939672.2939778`

[41] Ripke, S., et al.: Biological insights from 108 schizophrenia-associated genetic loci. Nature **511**(7510), 421–427 (2014)

[42] Schapke, J., et al.: EPGAT: Gene Essentiality Prediction With Graph Attention Networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics (2021)

[43] Shen, L., Thompson, P.M.: Brain Imaging Genomics: Integrated Analysis and Machine Learning. Proceedings of the IEEE **108** (jan 2020)

[44] Srinivasagopalan, S., et al.: A deep learning approach for diagnosing schizophrenic patients. Journal of Experimental and Theoretical Artificial Intelligence **31** (nov 2019)

[45] Stokes, J.M., et al.: A Deep Learning Approach to Antibiotic Discovery. Cell **180** (feb 2020)

[46] Van Der Maaten, L., et al.: Visualizing Data using t-SNE. In: Journal of Machine Learning Research. vol. 9 (2008)

[47] Van Kesteren, C.F., et al.: Immune involvement in the pathogenesis of schizophrenia: A meta-analysis on postmortem brain studies. Translational Psychiatry **7** (2017)

[48] Velicković, P., et al.: Graph attention networks. arXiv (oct 2017)

[49] Wang, et al.: Identifying Quantitative Trait Loci via Group-sparse Multitask Regression and Feature Selection: An Imaging Genetics Study of The ADNI Cohort. Bioinformatics (Oxford, England) **28**(2), 229–37 (2012)

[50] Wong, E.S., et al.: Interplay of cis and trans mechanisms driving transcription factor binding and gene expression evolution. Nature Communications **8** (2017)

[51] Yarkoni, T., et al.: Large-scale automated synthesis of human functional neuroimaging data. Nature Methods **8**(8), 665–670 (2011)

[52] Ying, R., You, J., Morris, C., Ren, X., Hamilton, W.L., Leskovec, J.: Hierarchical Graph Representation Learning with Differentiable Pooling. Advances in Neural Information Processing Systems (2018)

[53] Yuan, Y., Bar-Joseph, Z.: GCNG: graph convolutional networks for inferring gene interaction from spatial transcriptomics data. Genome biology (dec 2020)

[54] Zeng, L.L., et al.: Multi-Site Diagnostic Classification of Schizophrenia Using Discriminant Deep Learning with Functional Connectivity MRI. EBioMedicine **30** (apr 2018)

[55] Zhang, J., et al.: Predicting Disease-related RNA Associations based on Graph Convolutional Attention Network. In: Proceedings - 2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2019. Institute of Electrical and Electronics Engineers Inc. (2019)