RCT: Relational Connectivity Transformer for Enhanced Prediction of Absolute and Residual Intelligence

Mohammad Arafat Hussain[1], Ellen Grant[1,2], and Yangming Ou[1,2]

[1] Boston Children's Hospital, Harvard Medical School, Boston, MA 02115, USA
[2] Department of Radiology, Harvard Medical School, Boston, MA 02115, USA
{Mohammad.Hussain, Yangming.Ou}@childrens.harvard.edu

Abstract. This paper introduces the Relational Connectivity Transformer (RCT), a novel Graph-Transformer model designed for predicting absolute and residual full-scale intelligence quotient (FSIQ), performance IQ (PIQ), and verbal IQ (VIQ) scores from resting-state functional magnetic resonance imaging (rs-fMRI) data. Early prediction of neurocognitive impairments via IQ scores may allow for timely intervention. To this end, our RCT model leverages a relation-learning strategy from paired sample data via a novel graph-based transformer framework. Through a comprehensive comparison with state-of-the-art approaches in a 5-fold cross-validation setup, our model demonstrated superior performance. Statistical analysis confirmed the significant improvement ($p < 0.05$) in FSIQ prediction, strengthening the efficacy of the proposed method. This work marks the first application of a Graph-Transformer in predicting IQ scores using rs-fMRI, introducing a novel learning strategy and contributing to the ongoing efforts to enhance the accuracy and reliability of human intelligence predictions based on functional brain connectivity. The code is available in this GitHub repository.[3]

Keywords: fMRI · Graph · Transformer · Neurocognition · Intelligence.

1 Introduction

Neurocognition encompasses memory, motor control, speech, information processing, comprehension, thinking, and reasoning [1]. Intelligence, representing mental abilities related to *neurocognition quality*, is influenced by factors like diseases, treatments, lifestyles, and environmental conditions [2]. Early identification of neurocognitive impairments provides a critical window for intervention, with childhood interventions improving outcomes and adult health gains, including enhanced survival, reduced complications, better quality of life, and lower treatment costs [3,4,5]. Assessing human intelligence aids in predicting neurocognitive outcomes, but current research explains less than 1/3 of the variability [6].

Understanding the neural basis of human intelligence is a key goal in cognitive neuroscience [7]. Human neurocognition heavily relies on brain structure and

[3] https://github.com/marafathussain/RelationalConnectivityTransformer

structural and functional connectivity. While we acknowledge the importance of the brain's structure and its connections, we focus on functional connectivity and investigate its role in predicting neurocognitive outcomes in this study. Prior studies using brain magnetic resonance imaging (MRI) have linked human intelligence to the structure and function of different brain regions [8]. Recent research emphasizes the crucial role of interactions within and between functional brain networks in explaining individual variations in intelligence, particularly observed during resting-state functional MRI (rs-fMRI) [9,10].

Research has examined the link between brain connectivity and intelligence quotient (IQ) (review [8]). Also, few studies have predicted neurocognitive scores utilizing brain connectivity. For instance, Shen et al. [11] used a connectome-based model to predict brain-behavior relationships. He et al. [12] compared deep neural networks and kernel regression for predicting behavioral scores. Qu et al. [13] used a gated graph transformer model for similar predictions. Hanik et al. [14] was the first to predict general intelligence, specifically full-scale IQ (FSIQ) and verbal IQ (VIQ), using a regression graph neural network (GNN) applied to rs-fMRI. However, they did not consider predicting residual IQ scores that avoid the influence of demographic and socio-economic factors.

Brain connectivity data are structured with nodes and edges. Kawahara et al. [15] introduced BrainNetCNN, a convolutional neural network (CNN) variant, to process this data, but CNNs have limitations in the local spatial hierarchy. To address this, GNN models were developed, representing rs-fMRI-based brain networks as graphs with ROIs as nodes and functional activation correlations as edges. FBNetGen [16] used rs-fMRI data to enable learnable brain network generation. Transformers, with their self-attention mechanisms capturing long-range dependencies, led to Graph-Transformers, bridging the gap between GNNs and Transformers [17,18]. Kan et al. [19] proposed a Graph-Transformer variant, BNT, for rs-fMRI-based disease and sex predictions. However, using Graph-Transformers for human intelligence prediction with functional brain connectivity remains unexplored.

Despite the advancements in predicting neurocognitive scores using brain connectivity, current state-of-the-art (SOTA) models have limitations. They often overlook the influence of demographic and socio-economic factors on IQ, and the local spatial hierarchy of CNNs or the limited context-capturing ability of traditional GNNs restricts their performance. Furthermore, the predictive power of Graph-Transformers for human intelligence using functional brain connectivity remains underexplored. Our paper addresses these gaps by introducing the Relational Connectivity Transformer (RCT). This novel Graph-Transformer model incorporates a pair-wise relation learning strategy to predict absolute and residual FSIQ, performance IQ (PIQ), and VIQ scores from rs-fMRI data. Notably, our major contributions are fourfold:

1. It introduces the first Graph-Transformer-based approach for predicting absolute and residual IQ scores using rs-fMRI, addressing gaps in the SOTA by (a) predicting neurocognitive scores with Graph-Transformers for the first time, and (b) predicting 'residual IQ' scores along with absolute IQ scores.

2. It pioneers pair-wise relation learning [20] for neurocognitive prediction, processing data from two randomly selected subjects to learn relative relationships (i.e., effectively using each other as references) and improve IQ prediction accuracy. This relational learning enhances performance compared to traditional SOTA methods.
3. It predicts four distinct relations (cumulative, relative, maximal, and minimal) to enhance task-specific learning.
4. The pair-wise input strategy increases the training sample size from n to ${}^{n}\mathcal{P}_{r}$ (where $\mathcal{P}$ denotes permutation, n is the training sample size, and $r = 2$), addressing the infeasibility of rs-fMRI data augmentation.

2 Materials and Methods

2.1 Data

We gathered 1,009 rs-fMRI data from the public Autism Brain Imaging Data Exchange (ABIDE) [21]. After excluding subjects with missing PIQ, VIQ, and FSIQ scores, $N = 809$ subjects remained (age: 6-64 years, mean: 16.63±7.26; male/female: 682/127; Autism Spectrum Disorder (ASD)/neurotypical (NT): 401/408) from 15 sites. The rs-fMRI data were preprocessed using the Configurable Pipeline for the Analysis of Connectomes (CPAC) with brain regions defined by the Craddock 200 atlas [22], resulting in 200 brain ROIs per sample. We then computed residual IQ scores (rFSIQ, rPIQ, rVIQ) for all subjects, considering age, sex, diagnostic group (ASD or NT), and data collection site as independent variables. The formula used was: $rT = T - (\alpha + \beta A + \gamma S + \delta D + \eta E)$, where rT and T represent residual and absolute IQ scores, respectively. The variables A, S, D, and E correspond to age, sex, diagnostic group, and sample collection site, respectively. α, β, γ, δ, and η are parameters of linear regression. A similar approach has been used for residual fluid intelligence estimation [23], however, unlike [23], our dataset lacks information on parental education and income. Absolute and residual IQ ranges are [41,180] and [-65,78], respectively.

2.2 Relational Connectivity Transformer (RCT)

Our proposed RCT model (Fig. 1) takes a pair of rs-fMRI inputs ($X_1 \rightarrow t_1$, $X_2 \rightarrow t_2$), where $X \in \mathcal{R}^{P \times P}$, P is the number of brain ROIs, and t_1 and t_2 are the ground-truth IQ scores for subjects 1 and 2, respectively. These inputs, X_1 and X_2 are functional correlation matrices for subjects 1 and 2, selected randomly from the data pool, processed in parallel through two backbone transformer modules. The outputs of backbone modules are subsequently concatenated and fed to a relational transformer module that predicts four target relations defined as (1) cumulative relation, $\hat{r}_1 = \hat{t}_1 + \hat{t}_2$, (2) relative relation, $\hat{r}_2 = \hat{t}_1 - \hat{t}_2$, (3) maximal relation, $\hat{r}_3 = \max(\hat{t}_1, \hat{t}_2)$, and (4) minimal relation, $\hat{r}_4 = \min(\hat{t}_1, \hat{t}_2)$.

Backbone Modules. Our backbone modules process inputs (X_1, X_2) in parallel through a L-layer Multi-Head Self Attention (MHSA) mechanisms followed by a concatenation defined as:

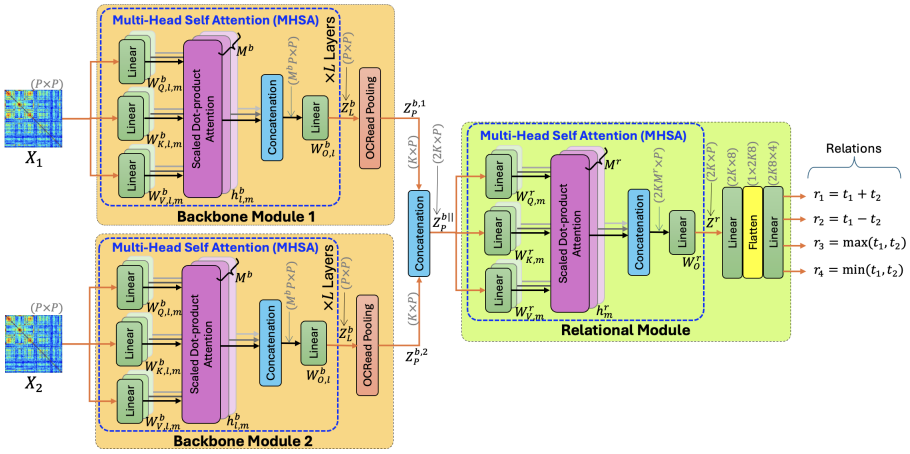

Fig. 1. The overall framework of our proposed Relational Connectivity Transformer.

$$Z_l^b = (\|_{m=1}^{M^b} h_{l,m}^b) W_{O,l}^b;\ h_{l,m}^b = \text{Softmax}\left(\frac{W_{Q,l,m}^b Z_{l-1}^b (W_{K,l,m}^b Z_{l-1}^b)^T}{\sqrt{d_{K,l,m}^b}}\right) W_{V,l,m}^b Z_{l-1}^b,$$

where $Z_0^b = X$ (omitting subscripts for inputs to avoid complexity), $\|$ denotes concatenation, M^b denotes the number of heads in the MHSA, $1 \leq l \leq L$ is the layer index, $W_{Q,l,m}^b$, $W_{K,l,m}^b$, $W_{V,l,m}^b$, and $W_{O,l}^b$ are learnable query, key, value, and condensation weight matrices, respectively, and $d_{K,l,m}^b$ is the first dimension of $W_{K,l,m}^b$ matrix. Typically, graph-based deep models require positional embedding for nodes [17]. For Graph-Transformers, eigen decomposition is used for positional embedding. However, it is shown in [24,19] that row indexes in a brain network adjacency matrix (rs-fMRI connectivity matrix X of size $P \times P = 200 \times 200$ in this study) can sufficiently provide positional information for each node, and therefore eigenvalue decomposition becomes redundant. Furthermore, in Graph-Transformer models, the MHSA mechanism typically requires combining node positions with the edge weights for attention calculation. However, it is also shown in [24] for brain networks that incorporating edge weights into the attention score calculation often degrades performance. Rather, pairwise dependency (pairwise correlations $X^{i,j}$ between blood-oxygen-level-dependent (BOLD) time courses of two ROIs i and j ($1 \leq i, j \leq 200$) in this study) is shown [19] sufficient for rs-fMRI-based Transformer study. The pooling function, also called readout, is a key component in graph representation learning. In this study, we adopted the state-of-the-art orthogonal clustering readout (OCRead) [19] in the backbone transformer modules (see Fig. 1). OCRead generates feature map Z_P^b by pooling from K clusters of functionally similar nodes in Z_L^b through orthonormal projection.

Relational Module. We concatenate the pooled features from backbone modules as $Z_P^{b\|} = Z_P^{b,1} \parallel Z_P^{b,2}$ (see Fig. 1). Afterward, our relational module processes the combined feature map $Z_P^{b\|}$ through a single layer MHSA mechanism

followed by a concatenation defined as:

$$Z^r = (\|_{m=1}^{M^r} h_m^r) W_O^r;\ h_m^r = \text{Softmax}\left(\frac{W_{Q,m}^r Z_P^{b\|} (W_{K,m}^r Z_P^{b\|})^T}{\sqrt{d_{K,m}^r}}\right) W_{V,m}^r Z_P^{b\|},$$

where M^r denotes the number of heads in the MHSA, $W_{Q,m}^r$, $W_{K,m}^r$, $W_{V,m}^r$, and W_O^r are learnable query, key, value, and condensation weight matrices, respectively, and $d_{K,m}^r$ is the first dimension of $W_{K,m}^r$ matrix. Finally, we flat the feature matrix Z^r and feed to a fully connected layer that generates four predictions $\hat{r}_1$, $\hat{r}_2$, $\hat{r}_3$, and $\hat{r}_4$. We use the mean square error (MSE) loss to train our RCT model defined as $\mathcal{L}_{MSE} = \frac{1}{4}\sum_{k=1}^{4}(r_k - \hat{r}_k)^2$, where actual target relations are defined as $r_1 = t_1 + t_2$, $r_1 = t_1 - t_2$, $r_3 = \max(t_1, t_2)$, and $r_4 = \min(t_1, t_2)$.

Implementation Details. In our backbone modules, we employ two MHSA layers ($L = 2$), and in the relational module, a single MHSA layer is used. The parameters $M^b = 4$ and $M^r = 8$ are set for the backbone and relational modules, respectively. OCRead pooling is conducted with $K = 10$ clusters, following the recommendation for the ABIDE rs-fMRI dataset in [19]. Xavier uniform initialization [25] initializes K orthonormal bases as cluster centers. The proposed method is evaluated using 5-fold cross-validation. For training, the Adam optimizer is employed with an initial learning rate of 0.0001, weight decay of 0.0001, and a batch size of 16. The original training sample size n is significantly increased through sample permutation ${}^n\mathcal{P}_2$ due to the input design. This augmentation exposes our model to numerous input combinations within a single epoch, leading to saturation of validation loss and accuracy within 10 epochs. Consequently, we set the epoch to 15. We also ensure that our training and validation data remain distinct throughout the permutation process. Our models are implemented in PyTorch version 1.12.1 and Python version 3.9, and training is conducted on an Intel E5-2650 v4 Broadwell 2.2 GHz processor, an Nvidia Titan RTX GPU with 24 GB of VRAM, and 16 GB of RAM.

3 Results

We estimate target IQ scores $\hat{t_1}$ and $\hat{t_2}$ from the predicted relations $\hat{r}_1$, $\hat{r}_2$, $\hat{r}_3$, and $\hat{r}_4$ for an input pair (X_1, X_2) as $\hat{t_1} = (\hat{r}_1 + \hat{r}_2)/2$; $\hat{t_2} = (\hat{r}_1 - \hat{r}_2)/2$; $\hat{t_1} = \hat{r}_3$ if $\hat{r}_2 > 0$ else $\hat{r}_4$; and $\hat{t_2} = \hat{r}_4$ if $\hat{r}_2 > 0$ else $\hat{r}_3$. Since there are two estimates of $\hat{t_1}$ and $\hat{t_2}$ for each input pair (X_1, X_2), we estimate the average of $\hat{t_1}$ pairs and $\hat{t_2}$ pairs as final predicted scores. For evaluating the prediction performance, we use mean absolute error (MAE) as $\frac{1}{V}\sum_{k=1}^{V}|q_k - \hat{q}_k|$, mean absolute percentage error (MAPE) as $\frac{1}{V}\sum_{k=1}^{V}(|q_k - \hat{q}_k|/|q_k|)$, and mean square error (MSE) as $\frac{1}{V}\sum_{k=1}^{V}(q_k - \hat{q}_k)^2$, where V is the total number of predictions during validation, q is the ground-truth IQ score, and $\hat{q}$ is the predicted IQ score.

3.1 Ablation Study

To choose the optimal configuration for our RCT, we tested four configurations (architectures 1-4) shown in Fig. 2(a)-(d). Architecture 1 has shared backbone

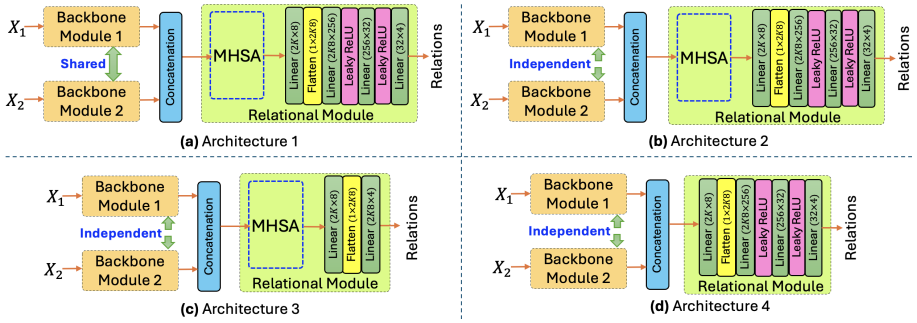

Fig. 2. Graph-Transformer architectures used in ablations studies: Architectures (a) 1, (b) 2, (c) 3, and (d) 4 (details in main texts). Colors are used in different blocks to correspond to respective colored blocks in Fig. 1.

Models	Architecture 1		Architecture 2		Architecture 3		Architecture 4	
Target	FSIQ	rFSIQ	FSIQ	rFSIQ	FSIQ	rFSIQ	FSIQ	rFSIQ
Fold 1	9.99±7.70	10.49±8.41	10.72±8.25	10.62±7.88	10.50±7.88	10.84±8.57	10.61±7.98	10.23±7.79
Fold 2	11.70±9.93	12.48±10.80	11.89±10.06	12.10±10.46	11.65±9.98	12.35±10.28	11.63±9.74	12.23±9.94
Fold 3	11.93±8.31	12.28±9.17	11.63±8.24	11.71±8.27	11.83±8.40	11.81±8.42	12.15±8.27	11.93±8.60
Fold 4	12.01±10.09	12.20±9.53	12.46±10.31	11.67±9.83	12.33±9.84	12.12±9.97	12.46±10.15	12.07±9.83
Fold 5	10.90±7.42	11.74±8.44	10.68±7.51	11.28±7.92	10.71±7.24	11.05±7.67	10.87±7.95	11.43±7.95
Target	PIQ	rPIQ	PIQ	rPIQ	PIQ	rPIQ	PIQ	rPIQ
Fold 1	12.17±9.10	11.81±9.25	11.78±8.92	11.85±8.73	11.64±8.93	12.16±9.26	12.62±9.39	12.08±9.04
Fold 2	12.32±10.65	12.44±10.89	12.30±10.82	13.04±10.92	12.30±10.59	12.80±10.99	12.14±10.49	12.86±11.09
Fold 3	12.36±9.90	13.01±10.60	12.76±9.45	12.49±9.53	12.23±8.97	12.76±9.64	12.63±9.48	12.55±9.38
Fold 4	13.26±10.19	12.14±9.18	12.38±9.80	11.82±9.64	12.39±9.70	12.06±9.52	12.21±9.62	12.23±9.37
Fold 5	10.80±8.79	11.31±8.82	10.63±9.02	11.65±8.71	10.75±8.69	11.02±8.93	10.85±8.99	10.94±8.65
Target	VIQ	rVIQ	VIQ	rVIQ	VIQ	rVIQ	VIQ	rVIQ
Fold 1	11.31±9.49	11.55±9.09	10.86±9.12	11.39±9.07	10.97±8.99	11.08±8.88	11.35±9.18	11.52±9.16
Fold 2	13.71±11.60	12.73±10.43	13.16±10.55	12.44±10.23	12.44±10.02	12.61±10.57	12.41±9.61	12.90±10.12
Fold 3	13.27±10.32	13.91±10.80	13.12±9.85	13.32±10.20	13.10±10.15	13.39±10.10	13.70±9.65	13.48±10.28
Fold 4	14.62±11.01	13.71±11.26	14.03±10.97	13.76±11.27	14.12±10.83	12.06±9.52	13.60±11.10	13.55±10.77
Fold 5	11.60±8.86	12.31±9.65	11.48±8.80	12.12±9.37	11.51±8.58	12.80±8.93	11.50±8.68	12.19±9.33
Mean	12.19		12.03		11.98		12.10	

Table 1. Absolute and residual IQ prediction performance in terms of MAE by four different architectures used in our ablation study (see Fig. 2).

modules (same learnable weights) for input pairs, with one MHSA and four fully connected layers (labeled as 'Linear') in the relational module. Architecture 2 has independent backbone modules (different learnable weights) for input pairs, with the same relational module as Architecture 1. Architecture 3 has independent backbone modules and a relational module with one MHSA and two fully connected layers. Architecture 4 has independent backbone modules and a relational module with no MHSA but four fully connected layers. We evaluated these architectures using 5-fold cross-validation, and the MAE results in Table 1 show that Architecture 3 has the lowest mean MAE. Thus, we selected Architecture 3 as the optimal configuration for our proposed RCT model (Fig. 1).

Justification of our Ablation Study Design. The ablation study in this work employs 5-fold cross-validation on the same dataset used for reporting results to determine the optimal deep transformer architecture among four different configurations. Cross-validation was chosen to maximize the robustness and generalizability of model performance by evaluating the model across multiple data splits, thereby reducing the risk of overfitting. Although an independent validation set is often preferred for hyperparameter tuning, cross-validation al-

lows the entire dataset to contribute to both training and validation, enhancing the reliability of performance metrics.

Insight into Architecture 3's Performance. Architecture 3 demonstrates the lowest mean MAE, likely due to its balanced complexity and ability to capture intricate patterns in the rs-fMRI data. The use of independent backbone modules enables the model to learn distinct features from input pairs, enhancing its capacity to represent complex relationships inherent in functional connectivity data. Additionally, the relational module's structure, consisting of one MHSA layer and two fully connected layers, provides an optimal balance between capturing global dependencies and maintaining sufficient model capacity without overfitting. This configuration likely allows Architecture 3 to effectively model the nuanced and high-dimensional nature of rs-fMRI data, leading to superior predictive performance.

3.2 Comparison to State-of-the-art

To compare absolute and residual IQ prediction performance by the proposed RCT method with state-of-the-art approaches, we implemented a CNN-based graph representation learning approach BrainNetCNN [15], a task-specific graph generation-based GNN approach FBNetGen [16], a conventional graph transformer approach Graphormer [17], and a rs-fMRI-tailored Graph-Transformer approach BNT [19]. We ran all these methods for 200 epochs in a 5-fold cross-validation setup. We used conventional single-input single-target ($X \rightarrow t$) feed-forward settings for these approaches and adhered to their authors-suggested hyperparameter settings. For the proposed approach, we tested two prediction scenarios: (1) both validation inputs are the same (i.e., from the same subject, $X_1 = X_2$), and (2) validation input pairs are mixed, i.e., X_1=X_2; $X_1 \neq X_2$. We present absolute and residual IQ prediction performance by different approaches in Tables 2 and 3, respectively, in terms of MAE, MAPE, and MSE. In both tables, we observe that the proposed RCT approach consistently exhibits the lowest absolute and residual FSIQ, PIQ, and VIQ prediction errors. However, it is noteworthy that the lowest MAE, MAPE, and MSE are not consistently associated with a specific input scenario; rather, the best performance varies between the scenarios $X_1 = X_2$ and $X_1 = X_2$; $X_1 \neq X_2$. In addition, we tested the statistical significance in prediction errors by the proposed RCT approach for $X_1 = X_2$ (to be fair with state-of-the-art) and by the best-performing state-of-the-art approach (i.e., FBNetGen in Table 2, and BrainNetCNN in Table 3). We found that the 2-tailed t-test on absolute error and absolute percentage error by the RCT and FBNetGen methods for the FSIQ score is significant for $p < 0.05$ (see Table 2). So, we can assume that the performance of the proposed RCT approach will also be statistically better than other approaches because other state-of-the-art performed worse than the FBNetGen method in absolute FSIQ prediction. This finding is also significant because the FSIQ score is a factor-weighted combination of the PIQ and VIQ scores [8].

Methods	Target	Metrics	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean
BrainNetCNN [15]	FSIQ	MAE ↓	18.55±14.10	20.85±15.20	22.43±14.15	20.55±14.80	21.17±15.01	20.71
		MAPE ↓	0.17±0.13	0.20±0.14	0.21±0.13	0.19±0.13	0.18±0.12	0.19
		MSE ↓	543.1±706.6	666.0±896.7	703.5±725.4	641.6±856.1	673.7±827.1	645.6
	PIQ	MAE ↓	19.43±14.45	21.14±15.54	21.76±14.23	19.77±15.12	20.93±14.95	20.61
		MAPE ↓	0.18±0.14	0.20±0.16	0.20±0.13	0.19±0.14	0.19±0.14	0.19
		MSE ↓	586.8±755.3	688.8±915.9	676.2±844.8	619.6±914.8	661.9±825.0	646.7
	VIQ	MAE ↓	19.44±14.22	21.46±16.21	22.89±15.55	20.88±15.71	21.66±16.38	21.27
		MAPE ↓	0.18±0.13	0.20±0.17	0.21±0.14	0.20±0.14	0.19±0.15	0.17
		MSE ↓	581.3±770.0	723.7±973.8	766.3±967.0	683.1±969.1	738.1±921.2	698.5
FBNetGen [16]	FSIQ	MAE ↓	11.78±8.55	12.21±9.74	12.56±8.51	12.43±10.38	12.41±8.16	12.28
		MAPE ↓	0.12±0.09	0.13±0.16	0.12±0.09	0.12±0.11	0.11±0.07	0.12
		MSE ↓	212.0±260.8	244.8±444.3	230.3±272.4	262.6±403.8	220.7±256.3	234.1
	PIQ	MAE ↓	12.58±9.44	**12.10±9.91**	12.96±9.63	12.34±9.63	11.91±8.61	12.38
		MAPE ↓	0.13±0.11	0.13±0.17	0.13±0.11	0.12±0.11	0.11±0.09	0.13
		MSE ↓	247.5±339.4	**244.6±459.5**	260.6±369.8	**245.1±372.5**	215.8±264.1	242.7
	VIQ	MAE ↓	12.19±9.40	12.56±10.68	13.58±9.06	**13.68±11.30**	12.93±8.98	12.99
		MAPE ↓	0.18±0.13	0.20±0.17	0.21±0.14	0.20±0.14	0.19±0.15	0.19
		MSE ↓	237.2±376.5	272.0±498.4	**266.4±315.9**	**314.9±530.8**	248.0±298.0	267.7
Graphormer [17]	FSIQ	MAE ↓	13.10±9.93	13.70±10.30	13.93±10.14	14.80±11.01	12.91±8.56	13.69
		MAPE ↓	0.13±0.11	0.14±0.15	0.13±0.10	0.15±0.12	0.12±0.08	0.14
		MSE ↓	270.3±380.4	294.0±465.6	296.8±380.0	340.3±448.8	239.8±276.6	288.2
	PIQ	MAE ↓	15.33±11.17	14.52±11.04	14.72±11.07	13.62±11.16	13.49±10.02	14.34
		MAPE ↓	0.16±0.14	0.15±0.17	0.15±0.13	0.14±0.13	0.13±0.11	0.15
		MSE ↓	359.9±519.0	332.8±518.4	339.5±482.5	310.0±468.8	282.5±384.1	324.9
	VIQ	MAE ↓	13.14±10.39	13.99±11.51	15.18±12.02	16.45±12.50	13.47±9.30	14.45
		MAPE ↓	0.13±0.14	0.14±0.16	0.15±0.13	0.17±0.15	0.12±0.08	0.14
		MSE ↓	280.7±446.6	328.4±528.7	375.3±576.3	427.1±687.8	267.9±371.1	335.9
BNT [19]	FSIQ	MAE ↓	11.96±9.52	12.80±10.41	13.49±8.60	13.59±10.35	11.89±8.15	12.75
		MAPE ↓	0.12±0.11	0.13±0.16	0.13±0.11	0.11±0.07	0.14±0.14	0.16
		MSE ↓	233.7±319.9	272.5±488.5	255.9±310.7	291.9±428.3	207.9±248.9	252.38
	PIQ	MAE ↓	13.70±10.83	13.28±11.48	13.75±10.67	13.03±11.25	12.77±9.75	13.31
		MAPE ↓	0.14±0.14	0.14±0.19	0.13±0.12	0.13±0.12	0.12±0.10	0.13
		MSE ↓	305.1±459.1	308.3±564.3	303.2±456.1	296.3±476.9	258.0±385.0	294.2
	VIQ	MAE ↓	12.11±9.68	13.93±10.83	14.48±10.77	15.11±12.35	12.57±9.76	13.64
		MAPE ↓	0.12±0.12	0.14±0.15	0.14±0.12	0.15±0.14	0.11±0.09	0.13
		MSE ↓	240.5±361.9	311.3±474.5	325.7±475.9	380.9±717.1	253.5±379.8	302.38
Proposed RCT (X_1=X_2)	FSIQ	MAE ↓	**10.37±7.75**	**11.55±9.89**	**11.77±8.36**	**12.24±9.79**	**10.63±7.19**	**11.31****
		MAPE ↓	**0.05±0.04**	**0.06±0.08**	**0.05±0.04**	**0.06±0.05**	**0.04±0.03**	**0.05****
		MSE ↓	172.9±238.5	**234.9±488.5**	**210.6±263.4**	249.3±359.8	167.4±197.7	**207.0**
	PIQ	MAE ↓	**11.59±8.92**	12.18±10.55	**12.22±8.89**	**12.20±9.69**	**10.69±8.63**	**11.77**
		MAPE ↓	**0.06±0.05**	**0.06±0.09**	**0.05±0.04**	**0.06±0.05**	**0.05±0.04**	**0.06**
		MSE ↓	215.8±307.2	262.7±583.4	230.6±329.8	247.8±346.1	191.3±293.1	**229.6**
	VIQ	MAE ↓	**10.77±8.93**	**12.39±9.96**	**12.94±10.14**	14.07±10.72	**11.49±8.49**	**12.33**
		MAPE ↓	**0.05±0.06**	**0.06±0.07**	**0.06±0.07**	**0.07±0.06**	**0.05±0.04**	**0.06**
		MSE ↓	201.6±346.2	**254.9±411.6**	274.9±438.0	317.6±535.5	206.6±297.5	251.1
Proposed RCT (X_1=X_2, $X_1 \neq X_2$)	FSIQ	MAE ↓	10.50±7.88	11.65±9.98	11.83±8.40	12.33±9.85	10.71±7.24	11.40
		MAPE ↓	0.10±0.09	0.12±0.16	0.11±0.08	0.12±0.10	0.09±0.06	0.12
		MSE ↓	**172.5±238.3**	235.7±490.9	210.7±263.2	**249.1±359.2**	**167.2±196.7**	**207.0**
	PIQ	MAE ↓	11.64±8.93	12.29±10.59	12.23±8.97	12.39±9.70	10.75±8.69	11.86
		MAPE ↓	0.12±0.11	0.13±0.19	0.11±0.10	0.12±0.11	0.10±0.09	0.12
		MSE ↓	**215.3±306.4**	263.5±584.6	**230.3±329.2**	247.6±345.5	**191.1±293.1**	**229.6**
	VIQ	MAE ↓	10.97±8.99	12.44±10.02	13.10±10.14	14.12±10.83	11.51±8.58	12.43
		MAPE ↓	0.11±0.11	0.12±0.14	0.13±0.11	0.14±0.13	0.10±0.08	0.12
		MSE ↓	**201.3±348.3**	255.2±411.4	274.7±437.0	316.8±535.5	**206.2±296.8**	**250.8**

Table 2. Comparison of absolute IQ prediction performance. **Bold red** fonts denote the lowest mean error per IQ category. The blue font represents the second-lowest mean error. **differences are significant for $p < 0.05$.

4 Discussion

Justification for Mixing ASD and NT data. In our study, we chose to combine data from both ASD and NT subjects for predicting IQ scores. This decision was driven by several factors: ***(1) Minimal Difference in IQ Scores***: The differences in actual IQ scores between ASD and NT subjects are relatively small compared to the typical error margins reported in SOTA (e.g., [14]) predictive models. As shown in Table 4, the absolute differences ($|\Delta|$) between means of different IQ types for ASD and NT groups are <7 points, whereas the MAE in SOTA models, including our results, is >9 points. This indicates that the variance between the actual IQs of ASD and NT subjects is less significant than the prediction error margin, making it less impactful in the context of our model's performance. ***(2) Aim to Improve Overall Predictive Accuracy***: The primary objective of our research is to reduce the gap between actual and predicted

Methods	Target	Metrics	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean
BrainNetCNN [15]	rFSIQ	MAE ↓	**10.32±7.63**	12.65±9.51	**11.70±7.84**	12.76±9.97	11.41±7.38	11.76
		MAPE ↓	2.35±6.49	6.00±7.05	4.01±3.32	4.15±23.05	2.02±0.25	3.71
		MSE ↓	**164.8±230.5**	326.2±395.4	**198.4±239.9**	337.8±376.2	184.7±221.1	242.4
	rPIQ	MAE ↓	**11.37±8.52**	13.75±9.93	**12.24±9.12**	12.01±9.17	12.13±7.83	12.30
		MAPE ↓	3.98±29.87	3.99±0.10	3.09±0.93	2.80±8.37	3.00±0.09	3.40
		MSE ↓	302.0±389.5	236.5±439.4	**233.2±323.7**	**228.5±319.7**	385.3±444.8	277.1
	rVIQ	MAE ↓	12.80±9.04	12.16±10.13	12.46±8.69	13.22±10.73	12.88±8.56	12.71
		MAPE ↓	3.80±4.63	**2.16±12.33**	3.08±3.68	4.42±23.67	3.00±3.09	3.29
		MSE ↓	**198.5±341.5**	250.5±428.3	231.1±275.3	290.1±508.6	214.6±281.3	237.0
FBNetGen [16]	rFSIQ	MAE ↓	10.41±8.18	12.86±9.63	12.72±7.96	12.38±9.70	11.81±9.57	12.04
		MAPE ↓	**2.06±9.29**	2.05±0.39	2.67±5.51	22.19±245.15	1.42±3.24	6.10
		MSE ↓	175.4±251.8	**233.5±391.8**	200.9±243.2	247.5±369.4	297.0±234.9	230.6
	rPIQ	MAE ↓	12.67±8.99	**12.00±9.66**	12.95±9.39	12.10±9.26	12.02±7.84	12.35
		MAPE ↓	4.54±16.55	2.39±2.65	2.98±4.29	3.14±3.59	3.06±3.50	3.22
		MSE ↓	**217.2±312.9**	**237.4±419.4**	246.3±345.1	232.2±324.5	383.0±440.5	263.2
	rVIQ	MAE ↓	12.85±9.26	13.28±10.04	12.57±8.74	13.30±11.03	12.01±8.75	12.80
		MAPE ↓	3.39±3.73	3.33±4.22	3.05±3.65	3.33±3.89	3.10±3.57	3.24
		MSE ↓	283.8±346.4	251.8±415.8	234.5±278.9	298.7±559.5	321.0±395.9	277.9
Graphormer [17]	rFSIQ	MAE ↓	11.43±8.44	12.64±10.28	13.49±9.10	13.23±10.34	12.91±8.63	12.74
		MAPE ↓	7.20±51.38	2.15±5.41	3.35±18.15	4.70±30.80	2.57±5.86	3.99
		MSE ↓	202.0±263.6	265.6±430.8	264.7±332.8	282.0±399.4	241.3±287.3	251.1
	rPIQ	MAE ↓	12.93±9.96	13.03±11.22	13.51±10.65	13.43±10.68	12.10±9.06	13.00
		MAPE ↓	2.80±11.97	3.67±8.12	3.70±2.36	4.98±41.85	2.23±4.29	3.48
		MSE ↓	266.6±377.1	295.9±555.1	296.1±480.7	294.5±425.5	228.7±425.5	276.4
	rVIQ	MAE ↓	11.75±8.85	13.68±10.87	14.63±10.54	14.22±11.92	13.61±10.61	13.58
		MAPE ↓	5.17±37.65	5.86±35.00	3.75±21.08	6.59±42.95	2.98±13.17	4.87
		MSE ↓	216.5±312.7	305.5±459.4	325.3±476.0	344.4±624.8	298.1±453.1	297.9
BNT [19]	rFSIQ	MAE ↓	11.49±7.99	12.98±10.06	12.23±8.35	12.74±9.72	11.17±8.10	12.12
		MAPE ↓	4.21±24.12	3.89±4.68	2.54±13.01	4.12±23.23	3.02±4.50	3.56
		MSE ↓	174.1±245.3	244.8±449.3	219.3±272.6	**232.5±357.0**	290.5±253.2	232.2
	rPIQ	MAE ↓	11.83±8.70	12.81±10.82	12.85±9.25	12.17±9.38	12.17±8.95	12.37
		MAPE ↓	3.85±3.88	2.95±3.66	3.65±2.56	2.43±10.94	3.68±2.44	3.31
		MSE ↓	275.7±304.2	281.5±556.4	250.6±344.3	236.3±345.5	285.0±321.5	265.8
	rVIQ	MAE ↓	11.49±8.68	12.67±10.74	13.53±9.95	13.34±10.96	12.31±9.10	12.67
		MAPE ↓	4.57±26.14	4.02±15.71	2.47±6.82	5.89±40.59	2.16±5.31	3.82
		MSE ↓	207.5±319.7	276.0±478.8	282.2±433.6	298.2±559.0	234.4±344.2	259.7
Proposed RCT (X_1=X_2)	rFSIQ	MAE ↓	10.60±8.38	**12.10±10.14**	11.75±8.38	**11.95±9.89**	**10.94±7.51**	**11.46**
		MAPE ↓	2.30±13.31	**0.99±2.46**	**1.18±4.63**	**2.97±22.24**	**0.94±2.55**	**1.67**
		MSE ↓	191.3±282.2	259.1±463.9	210.7±277.8	247.2±372.1	**181.1±232.4**	217.8
	rPIQ	MAE ↓	11.99±9.07	12.58±10.88	12.50±9.43	**11.91±9.41**	**10.84±8.70**	**11.96**
		MAPE ↓	**2.25±12.87**	**1.30±3.67**	**0.80±1.18**	**1.19±6.01**	**0.84±1.43**	**1.27**
		MSE ↓	234.5±357.6	285.3±569.5	255.8±352.3	237.0±349.5	**200.5±324.2**	**242.6**
	rVIQ	MAE ↓	**10.65±9.94**	**12.10±9.70**	**12.30±10.24**	13.07±11.72	**11.25±9.45**	**11.87**
		MAPE ↓	**2.55±16.14**	3.02±12.71	2.40±7.82	3.89±30.59	2.16±7.35	**2.80**
		MSE ↓	202.7±314.5	**204.9±390.6**	**224.7±408.1**	305.7±430.5	**201.5±290.2**	**227.9**
Proposed RCT (X_1=X_2, $X_1 \neq X_2$)	rFSIQ	MAE ↓	10.84±8.57	12.35±10.29	11.81±8.43	12.13±9.97	11.05±7.68	11.63
		MAPE ↓	4.73±32.55	2.17±5.32	2.47±10.11	6.10±44.45	2.13±8.47	3.52
		MSE ↓	191.1±281.3	258.5±465.4	210.7±277.5	246.6±370.7	181.3±233.2	**217.6**
	rPIQ	MAE ↓	12.16±9.26	12.80±11.01	12.76±9.64	12.06±9.53	11.02±8.93	12.16
		MAPE ↓	7.06±62.08	2.71±8.05	1.78±3.20	2.49±12.36	1.74±3.07	3.15
		MSE ↓	233.9±357.2	285.3±572.2	256.0±351.7	236.5±349.9	201.3±325.0	**242.6**
	rVIQ	MAE ↓	11.08±8.88	12.61±10.57	13.39±10.10	**12.06±9.52**	12.80±8.93	12.58
		MAPE ↓	3.39±14.92	4.68±24.51	**2.22±6.31**	**2.67±5.31**	**2.14±7.50**	3.02
		MSE ↓	201.7±329.3	271.0±464.5	281.5±444.6	**205.5±314.5**	203.2±291.5	232.5

Table 3. Comparison of residual IQ prediction performance. **Bold red** fonts denote the lowest mean error per IQ category. The blue font represents the second-lowest error.

IQ scores. By focusing on improving the overall predictive accuracy, regardless of the subject group, we aim to enhance the robustness and applicability of our model. The inclusion of both ASD and NT subjects helps in training a more generalized model that can potentially perform well across different populations. ***(3) Homogeneous Preprocessing and Residual Scores***: Our preprocessing pipeline and the calculation of residual IQ scores (rFSIQ, rPIQ, rVIQ) were designed to account for age, sex, diagnostic group (ASD or NT), and collection site as independent variables. This approach normalizes the data across these factors, ensuring that our predictions are not biased by group-specific variations and focusing on the inherent relationship between the brain connectivity features and IQ scores. ***(4) Relevance to Broader Applications***: Combining ASD and NT data allows our model to apply to a wider range of clinical and research settings where mixed populations are common. This enhances the practical utility of our findings, as models trained on mixed data are likely to be more adaptable and effective in diverse real-world scenarios.

	FSIQ	rFSIQ	PIQ	rPIQ	VIQ	rVIQ
ASD	105.92±16.10	0.067±15.40	105.87±16.63	0.007±15.99	105.58±16.75	-0.002±16.25
NT	111.00±12.19	0.005±11.71	108.32±13.00	0.009±12.50	112.01±12.88	0.041±12.54
$\|\Delta\|$	5.08	0.062	2.45	0.002	6.43	0.043

Table 4. Absolute differences ($|\Delta|$) between means of different IQ types for ASD and NT subject groups in the ABIDE dataset.

Novelty of the Proposed Approach. The key novelty of our approach can be summarized as: ***(1) Advancement Over Existing Methodologies***: The introduction of our RCT significantly advances existing methodologies, addressing unique challenges in predicting IQ scores from rs-fMRI data. ***(2) Innovative Relational Module Design***: While MHSA and linear layers have been used in other contexts, our relational module is designed to capture intricate pairwise relationships between brain regions across subjects, crucial for accurate cognitive predictions. Unlike traditional graph transformers that focus on node-level attention, our approach emphasizes relational dynamics between regions, drawing from relational modeling in neural networks [20]. ***(3) Novelty in Relational Learning Across Subjects***: Our model introduces relational learning across subjects, leveraging subject-to-subject variability to improve predictive performance and enhance model generalizability. ***(4) Task-Specific Learning with Distinct Relations***: Predicting distinct relations (cumulative, relative, maximal, and minimal) within brain connectivity enhances task-specific learning in neurocognition prediction, providing a richer understanding of the brain's functional architecture and improving predictive accuracy. ***(5) Pair-Wise Input Strategy and Increased Sample Size***: Our pair-wise input strategy enables training with a larger sample size (from n to ${}^{n}\mathcal{P}_2$, improving model robustness and performance by multiplying training examples in the rs-fMRI context. ***(6) Novel Application to IQ Prediction***: Although graph transformers have been applied to other rs-fMRI-based predictions like disease and sex classification [19], our work is the first to adapt this technology for IQ score prediction, filling a critical gap in the literature and demonstrating the versatility and robustness of graph transformers in a new domain. ***(7) Foundation for Future Research***: By focusing on IQ prediction, our study broadens the scope of graph transformer applications and provides a foundation for future research in neurocognitive assessments.

Limitations and Future Work. Our method shows promise but has limitations. The dataset is small with 809 subjects, though data permutation enhances sample size. We aim to increase diversity with more varied data from diverse backgrounds. The absence of some socioeconomic factors limits their inclusion in residual IQ estimation. Additionally, we focus solely on functional connectivity. Future research will integrate structural and diffusion MRI biomarkers for a more comprehensive prediction of neurocognitive outcomes.

5 Conclusion

We introduced the RCT, a novel Graph-Transformer model that leverages a pair-wise relation learning strategy for enhanced prediction of neurocognitive outcomes in terms of IQ scores. Our model demonstrated superior performance compared to SOTA approaches in a 5-fold cross-validation setup. The statistical significance (for $p < 0.05$) of the differences in prediction errors for absolute FSIQ strengthens the efficacy of the RCT approach. The incorporation of four different relations as targets further enhances the robustness of our task-specific learning. This work represents the first application of a Graph-Transformer for predicting neurocognitive outcomes using rs-fMRI and introduces a novel pair-wise relation learning strategy to the field, paving the way for more accurate and reliable predictions of human intelligence based on functional brain connectivity.

References

1. Watson, C.G., Stopp, C., Wypij, D., Bellinger, D.C., Newburger, J.W., Rivkin, M.J.: Altered white matter microstructure correlates with IQ and processing speed in children and adolescents post-fontan. The Journal of pediatrics **200** (2018) 140–149
2. Kessler, N., Feldmann, M., Schlosser, L., Rometsch, S., Brugger, P., Kottke, R., Knirsch, W., et al.: Structural brain abnormalities in adults with congenital heart disease: Prevalence and association with estimated intelligence quotient. International Journal of Cardiology **306** (2020) 61–66
3. Urschel, S., Bond, G.Y., Dinu, I.A., Moradi, F., Conway, J., Garcia-Guerra, G., et al.: Neurocognitive outcomes after heart transplantation in early childhood. The Journal of Heart and Lung Transplantation **37**(6) (2018) 740–748
4. Calderon, J., Bellinger, D.C.: Executive function deficits in congenital heart disease: why is intervention important? Cardiology in the Young **25**(7) (2015) 1238–1246
5. Beames, J.R., Kikas, K., Werner-Seidler, A.: Prevention and early intervention of depression in young people: an integrated narrative review of affective awareness and ecological momentary assessment. BMC psychology **9**(1) (2021) 113
6. Saha, S., Pagnozzi, A., Bradford, D., Fripp, J.: Predicting fluid intelligence in adolescence from structural MRI with deep learning methods. Intelligence **88** (2021) 101568
7. Hilger, K., Ekman, M., Fiebach, C.J., Basten, U.: Intelligence is associated with the modular structure of intrinsic brain networks. Scientific Reports **7**(1) (2017) 16088
8. Dizaji, A.S., Vieira, B.H., Khodaei, M.R., Ashrafi, M., Parham, E., Hosseinzadeh, G.A., Salmon, C.E.G., Soltanianzadeh, H.: Linking brain biology to intellectual endowment: A review on the associations of human intelligence with neuroimaging data. Basic and Clinical Neuroscience **12**(1) (2021) 1
9. Hilger, K., Ekman, M., Fiebach, C.J., Basten, U.: Efficient hubs in the intelligent brain: Nodal efficiency of hub regions in the salience network is associated with general intelligence. Intelligence **60** (2017) 10–25
10. Cole, M.W., Bassett, D.S., Power, J.D., Braver, T.S., Petersen, S.E.: Intrinsic and task-evoked network architectures of the human brain. Neuron **83**(1) (2014) 238–251

11. Shen, X., Finn, E.S., Scheinost, D., Rosenberg, M.D., Chun, M.M., Papademetris, X., Constable, R.T.: Using connectome-based predictive modeling to predict individual behavior from brain connectivity. nature Protocols **12**(3) (2017) 506–518
12. He, T., Kong, R., Holmes, A.J., Nguyen, M., Sabuncu, M.R., Eickhoff, S.B., Bzdok, D., Feng, J., Yeo, B.T.: Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. NeuroImage **206** (2020) 116276
13. Qu, G., Orlichenko, A., Wang, J., Zhang, G., Xiao, L., Zhang, K., Wilson, T.W., Stephen, J.M., Calhoun, V.D., Wang, Y.P.: Interpretable cognitive ability prediction: A comprehensive gated graph transformer framework for analyzing functional brain networks. IEEE Transactions on Medical Imaging (2023)
14. Hanik, M., Demirtaş, M.A., Gharsallaoui, M.A., Rekik, I.: Predicting cognitive scores with graph neural networks through sample selection learning. Brain Imaging and Behavior **16**(3) (2022) 1123–1138
15. Kawahara, J., Brown, C.J., Miller, S.P., Booth, B.G., Chau, V., Grunau, R.E., Zwicker, J.G., Hamarneh, G.: BrainNetCNN: Convolutional neural networks for brain networks; towards predicting neurodevelopment. NeuroImage **146** (2017) 1038–1049
16. Kan, X., Cui, H., Lukemire, J., Guo, Y., Yang, C.: Fbnetgen: Task-aware GNN-based fMRI analysis via functional brain network generation. In: International Conference on Medical Imaging with Deep Learning, PMLR (2022) 618–637
17. Ying, C., Cai, T., Luo, S., Zheng, S., Ke, G., He, D., Shen, Y., Liu, T.Y.: Do transformers really perform badly for graph representation? Advances in Neural Information Processing Systems **34** (2021) 28877–28888
18. Dwivedi, V., Bresson, X.: A generalization of transformer networks to graphs. arXiv preprint arXiv:2012.09699 (2020)
19. Kan, X., Dai, W., Cui, H., Zhang, Z., Guo, Y., Yang, C.: Brain network transformer. Advances in Neural Information Processing Systems **35** (2022) 25586–25599
20. He, S., Feng, Y., Grant, P.E., Ou, Y.: Deep relation learning for regression and its application to brain age estimation. IEEE Transactions on Medical Imaging **41**(9) (2022) 2304–2317
21. Di Martino, A., Yan, C.G., Li, Q., Denio, E., Castellanos, F.X., Alaerts, K., Anderson, J.S., Assaf, M., Bookheimer, S.Y., Dapretto, M., et al.: The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Molecular Psychiatry **19**(6) (2014) 659–667
22. Craddock, R.C., James, G.A., Holtzheimer III, P.E., Hu, X.P., Mayberg, H.S.: A whole brain fMRI atlas generated via spatially constrained spectral clustering. Human brain mapping **33**(8) (2012) 1914–1928
23. Pohl, K.M., Thompson, W.K., Adeli, E., Linguraru, M.G.: Adolescent brain cognitive development neurocognitive prediction. Lecture Notes in Computer Science, 1st edn. Springer, Cham (2019)
24. Cui, H., Dai, W., Zhu, Y., Kan, X., Gu, A.A.C., Lukemire, J., Zhan, L., He, L., Guo, Y., Yang, C.: Braingb: A benchmark for brain network analysis with graph neural networks. IEEE transactions on medical imaging **42**(2) (2022) 493–506
25. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics, JMLR Workshop and Conference Proceedings (2010) 249–256