Predicting Task Activation Maps from Resting-State Functional Connectivity using Deep Learning

Soren J. Madsen,[1*] Lucina Q. Uddin,[3] Jeanette A. Mumford[2], Deanna M. Barch[4], Damien A. Fair[5], Ian H. Gotlib[2], Russell A. Poldrack[2], Amy Kuceyeski[6], Manish Saggar[1*]

[1]Department of Psychiatry, Stanford University, USA

[2]Department of Psychology, Stanford University, USA

[3]Department of Psychiatry, University of California, Los Angeles, USA

[4]Department of Psychology, Washington University in St. Louis, USA

[5]Department of Pediatrics, University of Minnesota, USA

[6]Department of Radiology, Weill Cornell Medicine, USA

[*]Correspondence: sjmadsen@stanford.edu and saggar@stanford.edu

September 9, 2024

Keywords: deep learning, resting-state, task contrast, functional MRI

Abstract: Recent work has shown that deep learning is a powerful tool for predicting brain activation patterns evoked through various tasks using resting state features. We replicate and improve upon this recent work to introduce two models, BrainSERF and BrainSurfGCN, that perform at least as well as the state-of-the-art while greatly reducing memory and computational footprints. Our performance analysis observed that low predictability was associated with a possible lack of task engagement derived from behavioral performance. Furthermore, a deficiency in model performance was also observed for closely matched task contrasts, likely due to high individual variability confirmed by low test-retest reliability. Overall, we successfully replicate recently developed deep learning architecture and provide scalable models for further research.

1 Introduction

Task-based functional magnetic resonance imaging (tfMRI) remains a preferred neuroimaging modality to discover neural correlates (Emch et al., 2019; Ward et al., 2003) and examine the functional organization of the brain (Barch et al., 2013; Deary et al., 2004). Due to the high level of specificity of tasks performed in a scanner, experimental design can be fine-tuned to target specific neural circuits. However, the acquisition of tfMRI data can be highly time-consuming and expensive.

Recently, it's been suggested that task-based brain activity may contain more behaviorally relevant information than resting-state functional connectivity alone (Finn, 2021; Zhang et al., 2021). In addition, there are many clinical applications involving task performance and functional network analysis for psychiatric research, e.g., predicting responses to therapy (Pagliaccio et al., 2019) and analyzing individual differences in brain-behavior relationships (Hearne et al., 2021a; R. Jiang et al., 2020). As resting-state networks are known to contain similar functional architectures to those seen during task performance (Smith et al., 2009), the challenge of predicting task activation maps from resting-state features has become a more prominent focus in the field. Further, using resting-state data to predict task activation maps provides a cost-effective avenue for examining a variety of task contrasts rather than just a few and developing a proxy for task-based analysis of existing datasets that currently only contain resting-state data.

Models that can predict task activation maps from the resting state can be used to validate the Research Domain Criteria (RDoC) framework comprehensively. RDoC is a framework introduced by the National Institute of Mental Health to better understand mental disorders by focusing on dimensions of observable behavior and neurobiological measures. To validate this framework, task-based functional MRI contrasts provide a valuable approach by identifying overlapping patterns of brain activation across cognitive domains/sub-domains (Quah et al., 2024). However, given that no single individual is likely to perform all tasks associated with the RDoC domains, methods that can predict task contrast maps at the individual level are crucial. These predictive models would allow for a more comprehensive validation of the RDoC framework in healthy and patient populations. Recent work has shown that even clinically relevant dysfunctional activations could be accurately predicted from the resting state scans in patients with Schizophrenia (Hearne et al., 2021b).

Recent work by Cole et al., 2016 and Tavor et al., 2016 laid the foundational research for this idea by using predictors from resting-state networks (Yeo et al., 2011) in a series of regression models designed to predict task activation maps. More recently, this work was expanded upon by Ngo et al., 2022 to include a deep learning technique called BrainSurfCNN and suggest that the previous regression models like those in (Tavor et al., 2016) performed quite similarly to group-average contrast maps. To do this, Ngo et al., 2022 built upon several critical architectures in the field of deep learning like the U-Net (Milletari et al., 2016; Ronneberger et al., 2015) and surface-based mesh convolution (C. Jiang et al., 2019 along the cortical surface, represented by the fsLR_32k template, (Van Essen et al., 2012) to attain state-of-the-art performance.

In the current work, we aim to reproduce BrainSurfCNN and introduce two adjustments to the model: squeeze-excitation attention and graph convolution. Further, we elucidate several critical aspects of BrainSurfCNN's performance, including its sensitivity to the number of independent components and its overall efficacy in leveraging resting-state functional connectivity patterns for task activation mapping compared to our modified architectures. We also introduce two new models in accordance with these adjustments. In particular, we focus on our model, which maintains prediction performance while reducing computational costs. Lastly, as observed previously, models perform differently across subjects. To better understand the individual variability in model performance, we performed additional debugging. We specifically tested two hypotheses. First, we hypothesized that a lack of task engagement could drive the drop in predictability. We examined this hypothesis using behavioral performance on the task and its effect on model performance. Second, we hypothesized more noise in the scan would lead to poorer prediction performance. To test this, we analyze the temporal signal-to-noise ratio (tSNR) across the cortex and its overlap with the contrast maps.

2 Methods

2.1 Dataset

Like BrainSurfCNN, we used de-identified publically available data from the Human Connectome Project Young Adult (HCP) dataset (Van Essen et al., 2013) to train our models. The study was approved by the Stanford University Institutional Review Board.

Specifically, as input to our models, we used the FIX-cleaned, 3T resting-state fMRI (rsfMRI) data acquired in four 14.4-minute runs, each with 1200 time points per session per subject. We employ the data augmentation scheme as proposed in (Ngo et al., 2022) in which samples of the resting state data are drawn from all four recordings of resting state data. This aggregation of recording data amounts to 4,800 time points that are split into eight samples of 600 time points from which the functional connectivity is determined. These 8 samples allow us to extend the training set of our data by uniformly sampling from these 8 chunks in each epoch of training. The acquisition and preprocessing methods for the HCP dataset have been described elsewhere in (Barch et al., 2013; Glasser et al., 2013; Smith et al., 2013). Group-level parcellations derived from spatial ICA were also released by the HCP, and we examined the performance of our models trained on various component parcellations for computing the functional connectomes. Since the authors in (Ngo et al., 2022) explored data with 50 independent components (ICs), we thought a natural extension would be to explore various amounts of independent components around that order of magnitude. Using these spatial components, we create new training datasets derived from 15, 25, 50, and 100 independent components. For our target data, HCP's tfMRI data spans seven task domains, and, following Tavor et al., 2016 and Ngo et al., 2022, we obtained 47 unique contrasts after excluding redundant task contrasts.

Additionally, we only included subjects with all four resting-state runs and all 47 task contrasts, which gave us n=919 subjects for training, of which five subjects were used for a validation set (Ngo et al., 2022). The use of the validation set prevents overfitting our models during the training process. We also carried out a 5-fold cross-validation training regime and found similar performance in our models as seen in Supplementary Table 4. Similar to previous work (Ngo et al., 2022), our test set comprised n=39 subjects from the original HCP dataset that were retested with the same acquisition methods. This allowed us to compare the model's performance against a repeat contrast during a separate visit. The test set participants do not overlap with the participants in our training set. We only hold out these 39 participants because they have two sets of contrasts. We use this second set as an optimal prediction (though not model-based) of the activation seen in the ground truth task contrasts. The idea here is that a task contrast taken from the same subject is the best model in terms of predicting the individual variability seen in the same task in a different session: the same brain should look most similarly to the same brain across multiple scans. Thus, this retest provides us with a theoretical upper bound, or noise ceiling, for model performance. We differentiate the data we tested on from this noise ceiling benchmark by hereon referring to the test data as the "test dataset" and the noise ceiling benchmark as the "retest dataset." This is an important distinction because there is no resting-state data collected for the retest dataset which means, therefore we don't use this retest dataset as a ground truth for prediction.

Similar to Ngo et al., 2022, in the proposed deep learning framework, the input data are structured as multi-channel fs_LR polyhedral meshes, each consisting of 32,492 vertices, to accommodate the high-dimensional spatial topology of cerebral cortex representations. Each channel within these meshes is dedicated to an independent component derived from functional MRI data preprocessing, serving as a distinct feature for the model's input layer. For instance, in a scenario where the model is configured to analyze 50 independent components, the total input dimensionality would amount to 100 channels, reflecting the bilateral nature of cerebral anatomy with a separate mesh for each hemisphere. The model's output mirrors the input's structural format, generating a polyhedral mesh with 32,492 vertices. However, in the output mesh, each vertex's value across the channels indicates the predicted task contrast for the corresponding location in the brain hemisphere. Given that the study focuses on predicting 47 distinct task contrasts, the model's output dimensionality is adjusted to feature 94 channels, with each pair of channels representing the predicted contrasts for the left and right hemispheres, respectively.

2.2 Models

2.2.1 BrainSurfCNN

The development of BrainSurfCNN integrates the foundational principles of a U-Net architecture (Milletari et al., 2016; Ronneberger et al., 2015), a prevalent framework in medical image segmentation, with advanced mesh convolution techniques to facilitate the analysis of brain surface data. This innovative approach is significantly influenced by the pioneering work on convolution for spherical meshes called UGSCNN, as detailed by C. Jiang et al., 2019. The primary objective of BrainSurfCNN (Ngo et al., 2022) is to leverage the spatial hierarchies inherent in polyhedral meshes to predict task-related activation patterns, represented as contrast maps, from resting-state independent component (IC) maps.

In our study, we replicated the BrainSurfCNN model using the source code made publicly available by the original authors on GitHub [1]. This replication process was not merely an exercise in model reconstruction, but a deliberate effort to validate and extend the model's application. A cornerstone of our analysis involved a systematic examination of BrainSurfCNN's performance across a spectrum of resting-state scans characterized by varying numbers of independent components. Specifically, we conducted comparative analyses using datasets composed of 15, 25, 50, and 100 independent components to assess the model's robustness and scalability with varying levels of information. This approach allowed us to explore the impact of independent components on the model's ability to predict task contrasts accurately on both an individual and a group level of analysis.

2.2.2 BrainSERF

The Squeeze-and-Excitation (SE) attention mechanism has emerged as a pivotal innovation in the field of computer vision, significantly enhancing the representational power of convolutional neural networks (CNNs) by enabling channel-wise feature recalibration (Hu et al., 2018). This technique introduces trainable scaling coefficients that adaptively adjust the weighting of each channel in the network's feature maps, thereby allowing the model to emphasize informative features and suppress less useful ones

[1] GitHub Link: https://github.com/ngohgia/brain-surf-cnn

dynamically. The SE attention mechanism operates by first 'squeezing' global spatial information into a channel descriptor through global average pooling, followed by 'excitation' operations—fully connected layers that capture channel-wise dependencies. Given two weight matrices W_{sq} and W_{ex}, we perform the squeeze operation by globally pooling each mesh so that our representation of input X transforms from shape $\mathbb{R}^{32,492 \times C}$ to $\mathbb{R}^{1 \times C}$ where C represents the number of hidden channels. Next, we define a ratio r such that the matrix W_{sq} transforms the data from $\mathbb{R}^{1 \times C}$ to $\mathbb{R}^{1 \times \frac{C}{r}}$. We finish the 'squeezing' with a Tanh activation to capture negative values to get X'. To perform the 'excitation' step, we multiply the transformed X' by W_{ex} to expand the channel-wise axis back to a matrix in $\mathbb{R}^{1 \times C}$ to get S. We apply a Tanh activation function again to capture negative values. This matrix S represents a scaling on the channel-wise axis. We channel-wise multiply $X \in \mathbb{R}^{32,492 \times C}$ by S to achieve the output X_{scaled}, a channel-wise rescaled form of the input X. The intuition with this mechanism is to periodically rescale the hidden representation of the mesh channels and highlight or suppress entire latent activation patterns.

These operations produce scaling coefficients applied to the original feature maps, effectively allowing the network to perform self-attention across channels. In our network architecture, we implement SE attention before the mesh convolutional layers during the coarsening stage of the mesh, inspired by the work in Wang et al., 2021.

Furthermore, we have innovated beyond traditional activation functions by transitioning from Rectified Linear Unit (ReLU) activations to Hyperbolic Tangent (Tanh) activations throughout our network. This adjustment is primarily motivated by the need to accommodate negative values within the SE attention mechanism and the predicted task contrasts. The Tanh activation function enables our network to capture a broader spectrum of feature dynamics, including both positive and negative activations that are present in the task contrast maps.

2.3 BrainSurfGCN

In predicting task contrast activation maps from resting-state functional magnetic resonance imaging (rsfMRI), employing graph neural networks (GNNs) offers several compelling advantages that align with the intrinsic properties of brain data and the objectives of neuroimaging analysis. The human brain can be conceptualized as a complex network, with nodes representing different brain regions and edges representing functional or structural connections between these regions. This network-centric view is inherently compatible with the structure of GNNs. We define a graph below.

Definition 1. *An undirected graph G can be defined as $G = (\mathcal{V}, \mathcal{E})$, where $v \in \mathcal{V}$ represents a node with feature size in $\mathbb{R}^C$ and $e \in \mathcal{E}$ represents an undirected edge between two vertices v_i and v_j.*

We leverage the spherical mesh used in Ngo et al., 2022 as the graph's structure for each input, such that $|\mathcal{V}| =$ 32,492 mesh. Thus, our input data has two parts. The node-wise representation, $\mathcal{V}$, is represented as a matrix in $\mathbb{R}^{|\mathcal{V}|\times\mathcal{C}}$, where each node $v \in \mathcal{V}$ has C learnable parameters. The edge-wise representation is a list of tuples of vertices that are connected such that edge $e \in \mathcal{E}$ connecting vertices v_i and v_j is represented as (i, j), and $\mathcal{E} \in \mathbb{Z}^{2\times|\mathcal{E}|}$.

We build BrainSurfGCN upon the work of Kipf and Welling, 2016, which introduces the Graph Convolution Network. The layer architecture we use for this study builds upon the implementation of the Graph Convolution Layer seen in PyTorch Geometric (Fey and Lenssen, 2019), and we build a network composed of what we call the BDLayer. The structure of the BDLayer includes a GCN Layer, a LeakyReLU activation, and a BatchNorm layer. The BatchNorm layer accelerates learning by reducing internal covariate shift (Ioffe and Szegedy, 2015) through the recentering of features using the formula:

$$y = \frac{x - \mathbb{E}[x]}{\sqrt{\text{Var}[x]}} * \gamma + \beta \tag{1}$$

where γ and β are learnable parameters, this layer acts along a node's feature-wise axis. The GCN Layer defines the update of a node's feature space from X to X' as

$$X' = \hat{D}^{-1/2}\hat{A}\hat{D}^{-1/2}\Theta \tag{2}$$

where $\hat{A} = A + I$ denotes an adjacency matrix with added self-loops, and Θ denotes the learnable weight matrix. The adjacency matrix A is formed such that $A_{ij} = 1$ if $e_{ij} \in \mathcal{E}$, and $A_{ij} = 0$ if $e_{ij} \notin \mathcal{E}$. This operation allows us to pass information between vertices on the mesh at a distance of 1 hop per layer. To ensure that node information is disseminated across the entire graph, we use 8 BDLayers. Additionally, we believe that spatial information is important in predicting the activation on the mesh, so we included the XYZ coordinates of each vertex as 3 additional channels for our input data. These coordinates represent the locations of the vertices of the polyhedral sphere centered on the origin (0,0,0) with a radius of 1. Work in Liu et al., 2018 also suggests that including spatial information can improve convolution operations. With a resting-state map of 50 ICs per hemisphere, our input feature space lies in $\mathbb{R}^{32492\times(2*50+3)}$. Figure 1 shows our architecture from a high-level perspective, complete with the input and output feature spaces.

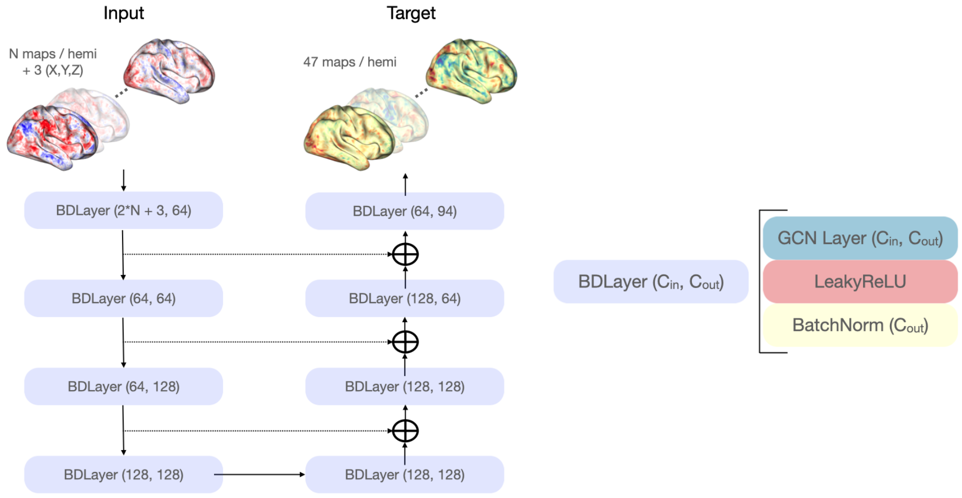

Figure 1: **A high-level overview of the BrainSurfGCN model architecture.** We chose this structure for the BDLayers of the model since the literature has shown that the combination of Graph Convolution, LeakyReLU, and BatchNorm performs well. There are 8 modules of BDLayers to ensure that information is disseminated across the entire mesh during one pass of the model.

2.4 Training Details

All the models trained in these experiments used the same set of hyperparameters and training regimes. Models were developed and trained in PyTorch (Paszke et al., 2019), and we used Adam (Kingma and Ba, 2014) for optimization. We utilized a learning rate of 0.001 and trained the models for 50 epochs using mean squared error (MSE) loss to promote task contrast reconstruction. We used the model with the best-predicted correlation on the validation set at the end of 50 epochs for evaluation. Because we trained several types of models, we compared performance after the first round of training to decide if training should continue. Two models did not continue after the first round of training, and they are mentioned in Sections A.1 and A.2 of the Supplementary Material.

If the model's performance was similar – measured by the correlation between prediction and ground truth – to the model performance demonstrated by Ngo et al., 2022, we continued with the next phase of training. In this second step, we used the trained model to compute and average the MSE across the training set for same-subject reconstruction loss, α, and across-subject reconstruction loss, γ, to be used as hyperparameters in the next step of training that used a reconstructive-contrastive (RC) loss function proposed by Ngo et al., 2022. Given a mini-batch of N samples, $B = \{\hat{x}_1..\hat{x}_N\}$, in which $\hat{x}_i$ is the target multi-channel contrast image of subject i, the RC loss function is defined as

$$L_R = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2 \text{ and } L_C = \frac{1}{\frac{1}{2}(N^2 - N)}\sum_{\hat{x}_j \in B, j \neq i}^{N}(x_i - \hat{x}_j)^2 \tag{3}$$

$$L_{RC} = [L_R - \alpha]_+ + [L_R - L_C + \gamma]_+. \tag{4}$$

Using the RC loss described in Equation 4, we train the model again for 50 epochs. Before training, we initialize values for α and γ from the losses computed over the training set. These values are updated every 10 epochs such that:

$$\alpha_t = \frac{1}{2}\alpha_{t-1} \text{ and } \gamma_t = 2\gamma_{t-1} \tag{5}$$

until α reaches a minimum of 1 and γ reaches a maximum of 10. Using this RC loss function, we encouraged the model to differentiate task contrast reconstructions between subjects, thereby capturing individual differences.

2.5 Evaluation

To evaluate the efficacy of our model in predicting task contrast maps from resting-state functional MRI (fMRI) data, we employ three distinct metrics: correlation, Dice coefficient (also referred to as F1 Score), and subject identification accuracy. These metrics offer a comprehensive assessment of the model's performance, each addressing different aspects of prediction quality.

The correlation metric primarily measures the linear relationship between the model's predicted task contrasts and the empirically observed (ground truth) contrasts. This metric was chosen for the sake of replication to compare the results from (Ngo et al., 2022. By computing Pearson's correlation coefficient, we quantify how the predicted values co-vary with the actual task contrasts across the mesh's vertices. A high correlation coefficient indicates that the model effectively captures the spatial distribution of neural activations associated with specific tasks, mirroring the patterns observed in the ground truth data. A crucial aspect of the performance is ensuring that within-subject covariance is more similar than inter-subject covariance.

The Dice coefficient enhances our evaluation by measuring spatial overlap between the predicted and ground truth task contrast activation patterns. To calculate this metric, we first apply a series of thresholds to the activation values at each vertex, classifying them as 'activated' or 'non-activated.' For each threshold, the Dice coefficient is computed as the harmonic mean of precision and recall between the two binary sets of activated vertices. Thus, the specificity of predicted activation on the contrasts is computed over a range of threshold values, ensuring predicted regions with higher activation (from a higher threshold) overlap with the higher activations in the ground truth contrasts. Thereby, we capture another measurement, in addition to correlation, related to both the spatial distribution and the relative magnitude of the cortical activation.

To assess the model's capability in capturing individual-specific features within the task contrasts, we use an evaluation method introduced by Ngo et al., 2022 based on subject identification accuracy. This approach involves computing the correlation between each predicted task contrast and all available ground truth contrasts across subjects for the same task. The identification of the subject is deemed correct if the highest correlation corresponds to the correct subject's ground truth contrast compared to those for other subjects. This metric reflects the model's sensitivity to individual differences in brain activation patterns, highlighting its potential for personalized neuroimaging analysis. We computed this accuracy for each quantity of ICs used in training and also included a bagged average across the models trained on various ICs to give us an average prediction. Unfortunately, GPU memory restrictions prevented us from training a version of BrainSERF on input data with 100 ICs.

3 Results

To provide a clear and unbiased comparison of model performance, we selected models with the highest subject identification accuracy for further analysis. These best-performing models were selected from the sets of models trained on the datasets constructed from selecting various independent components. For example, our experiments showed that BrainSurfCNN performed best with 25 ICs, while BrainSurfGCN performed best with 50 ICs. These models were selected for further analysis. This selection criteria ensures that our comparisons are centered around the top-performing models across all evaluated architectures.

Initially, we explored the spatial correlation of the predicted task activation maps and the subject identification accuracy. The subject identification was performed by verifying if a predicted task contrast had the highest correlation with the target contrast. This analysis is crucial for understanding how well the models can predict individually localized brain regions associated with specific cognitive tasks. These results for spatial correlation are shown in Figure 2. In this figure, a high correlation difference value signifies that the spatial correlation of a given prediction is substantially higher when comparing the prediction to the ground truth contrast of the same subject, as opposed to the ground truth contrasts of different subjects. Thus, we ensure the model interprets the idiosyncrasies of an individual's resting-state when making a prediction. Using the spatial correlation of a single predicted contrast with the ground truth contrasts, we attempt to classify which subject the predicted contrast derived from and call this subject identification accuracy. The averaged values across all task contrasts are shown in Table 1.

3.1 Model Evaluation

All models performed similarly and occasionally identified subjects more accurately than the retest contrast maps in certain tasks such as Language: Math and Language: Story. The readers are directed to Supplementary Figures 11 and 12 to see a more detailed view of model performance on each task contrast. To explore the possibility of a bagged model architecture, we average the predicted task contrasts across the models trained on various ICs and report their subject identification accuracy. The retest dataset contains a task contrasts from a second session of scans for our test set population. We use the retest dataset correlated with the target contrasts to determine a benchmark for subject identification accuracy, since the retest reliability provides a theoretical upper bound on model performance. When we computed the same metrics across this retest set, the retest contrasts scored an average accuracy of 92.9% across all 47 tasks. A more detailed visual of subject identification accuracy across the best-performing individual models is shown in Supplementary Figure 4.

Table 1: Subject Identification Test Accuracy Across Varying ICs

Model	15 ICs	25 ICs	50 ICs	100 ICs	Avg
BrainSurfCNN	80.3	81.0	79.5	79.3	81.8
BrainSERF (ours)	**80.6**	**82.3**	80.4	N/A	**82.8**
BrainSurfGCN (ours)	79.5	79.5	**80.7**	**79.9**	82.4
Retest (Benchmark)	**92.9**				

Table 2: Number of Trainable Parameters for Each Model

No. Independent Components	BrainSurfCNN	BrainSERF	BrainSurfGCN
15 ICs	359,902	363,062	**23,742**
25 ICs	365,022	368,846	**24,382**
50 ICs	377,822	384,706	**25,982**
100 ICs	403,422	422,426	**29,182**

Additionally, we report the number of parameters for each model with the varying number of ICS to emphasize BrainSurfGCN's drastic reduction in size. By utilizing shared parameters across all nodes,

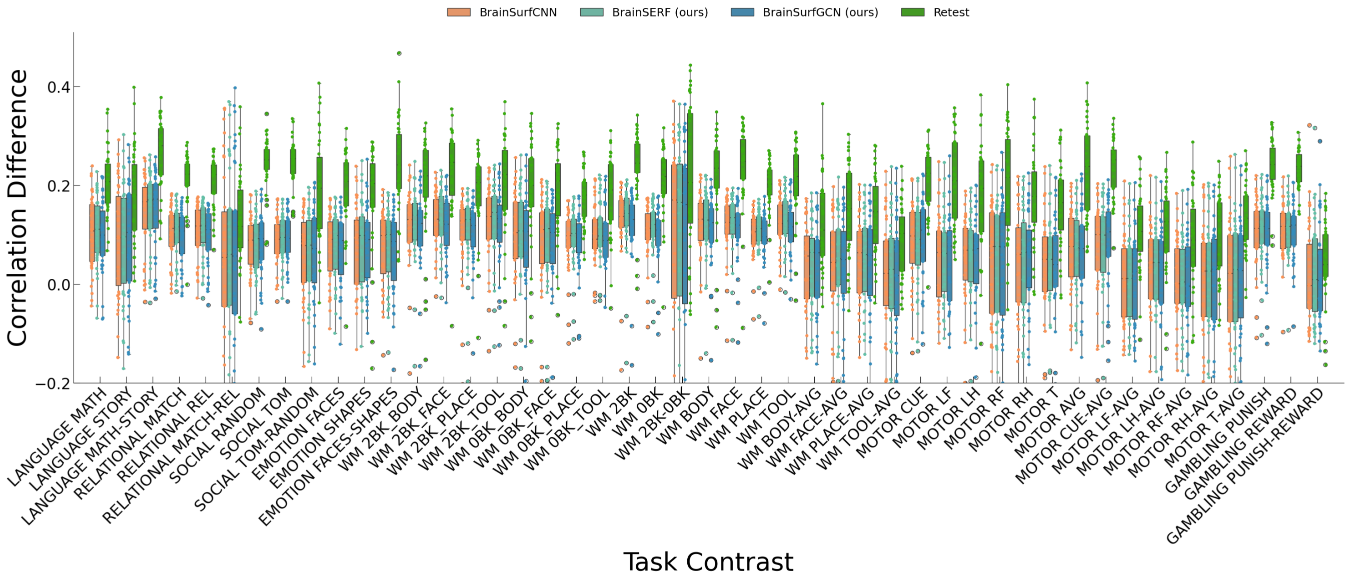

Figure 2: **Difference in spatial correlation of predicted contrast to ground truth contrast of the same subject versus the mean spatial correlation of predicted contrast to ground truth contrasts of all other test set subjects.** The correlation metric primarily measures the linear relationship between the model's predicted task contrasts and the empirically observed (ground truth) contrasts. In this analysis, we derive the metric displayed in the figure by first calculating the correlation between a prediction based on a single subject's resting-state data and the subject's ground truth task activation contrast. Subsequently, we subtract the mean correlation of the same prediction from the task activation contrasts of all other subjects. A high correlation difference value signifies that the spatial correlation of a given prediction is substantially higher when comparing the prediction to the ground truth contrast of the same subject, as opposed to the ground truth contrasts of different subjects. This indicates that the model proficiently captures the idiosyncratic features of each individual's task-related brain activity.

Table 3: GPU Usage (MB) for Each Model

No. Independent Components	BrainSurfCNN	BrainSERF	BrainSurfGCN
15 ICs	111.460864	111.476224	**0.115712**
25 ICs	111.481344	111.498752	**0.118272**
50 ICs	111.532544	111.562240	**0.124416**
100 ICs	111.634944	111.713280	**0.137216**

BrainSurfGCN efficiently leverages the structure of a mesh without the necessity of storing the mesh gradients as a buffer. It is important to note that the numbers reported in Table 2 reflect the number of trainable parameters in the model, and the GPU memory required to store model parameters is reported in Table 3. These numbers tend not to change much as we adjust the number of ICs because the model's hidden states don't change at all. Operations over the number of ICs quickly become operations over a set of model parameters with predefined sizes.

To achieve a granular view of all models' predictive capabilities, we plotted the activation maps post-thresholding the most significantly activated vertices. Figure 3 shows activation maps for a representative participant (id: 917255) prepared for three distinct threshold values—10%, 25%, and 50% during the Social Cognition: Theory of Mind task. The models' predictions varied with different levels of thresholding, offering insights into their sensitivity and specificity at various levels of whole-brain activation.

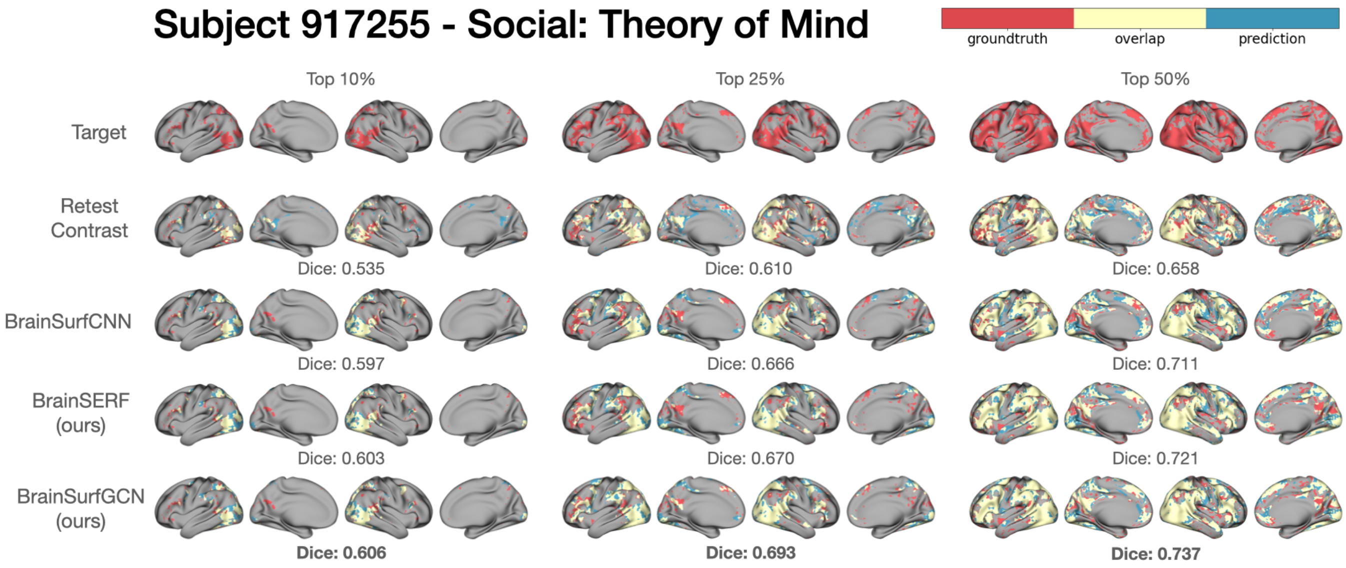

Figure 3: **Thresholded task activation of subject 917255 for Social Cognition: Theory of Mind.** We threshold the activation maps at 3 different values to show how effectively the models capture the distribution of activation in the brain. We also show the Dice overlap score to underscore performance improvements of BrainSERF and BrainSurfGCN that can be seen on the individual level.

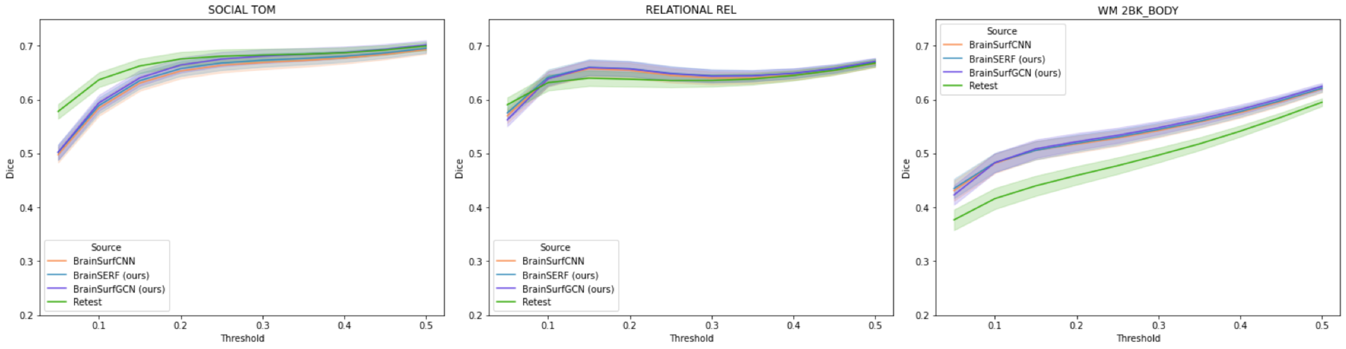

Figure 4: **Dice score plots for three selected tasks showing variations in performance across task contrasts.** Each of these plots shows a variation in model performance when compared to the retest. The first shows asymptotic convergence to retest contrast performance. The second shows similar performance between model predictions and retest contrasts, and the third plot shows improved performance over the retest contrasts.

We computed Dice scores across a spectrum of threshold values, ranging from 5% to 50%, in increments of 5%, as originally proposed by the creators of BrainSurfCNN. This detailed analysis allowed us to plot the Dice scores' variation with changing thresholds, offering a nuanced understanding of model performance across a broader spectrum of activation intensities as seen in Figure 4. In this figure, we also noticed that the model predictions seem to outperform the theoretical upper bound or noise ceiling. In this case, the characteristics of real data from the Retest contrasts mean that noise or small shifts in the location of activation could affect the performance here. These results are consistent with what Ngo et al., 2022 reported.

Furthermore, we calculated the area under the curve (AUC) for these Dice scores, which serves as a comprehensive measure of model performance across all thresholds. The unthresholded maps for all models were also examined to provide a baseline comparison of their predictive capabilities. The maximum value of the AUC is 0.45 since this number is the result of integrating a perfect Dice score over the range of threshold percentages. These values for the Dice AUC can be seen in Figure 5. Additionally, we included the average Dice AUC over the test set for all task contrasts in Figure 12 of the Supplementary Material.

For a more detailed exposition of the Dice AUC scores and additional plots of Dice scores across various tasks and models, readers are directed to Supplementary Tables 5 and 6. This extended analysis enriches our understanding of each model's strengths and weaknesses, facilitating a comprehensive evaluation beyond the scope of the main manuscript.

3.2 Debugging Performance

The model predictions and retest contrasts seem to follow similar trends in subject identification accuracy and spatial correlation when compared to the ground truth, and the variability in performance encouraged us to explore the reasons behind this phenomenon. We tested two hypotheses to explain this variability in performance. The first hypothesis is that a lack of neural engagement could drive the drop in predictability. Our goal was to determine if below-average task performance is related to a lack of focus on the task while in the scanner. A lack of focus may indicate a lack of neural engagement; hence, brain regions activated during the task would be different than expected, leading to lower predictability. To test this hypothesis, we identified the task contrasts that included an accuracy and a reaction time metric. For each of these tasks, we split the group of test subjects into above-average and below-average performers based on whether they performed above or below the mean accuracy or reaction time. Next, we plot these two groups along their correlation difference metric gathered from predictions made by

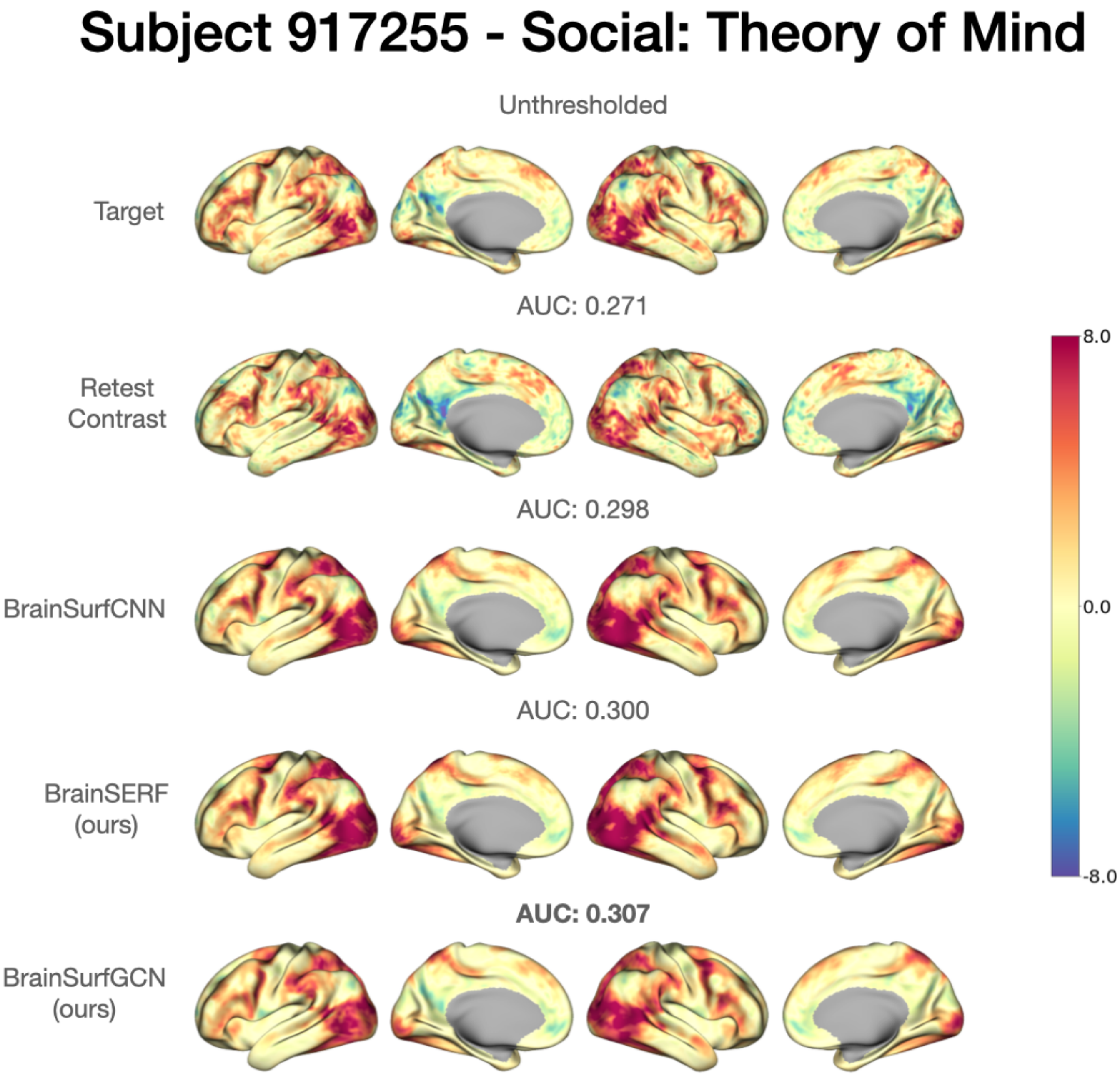

Figure 5: **Unthresholded task activation of subject 917255 for Social Cognition: Theory of Mind** We compare ground truth, retest, and model predicted activation for an individual. Model predictions are much smoother on the whole than the ground truth or retest activations. Though smoother, the models perform higher in the Dice AUC score than the retest contrasts.

BrainSurfGCN.

In Figure 6, we identified the difference between above-average and below-average groups as statistically significant for the retest and predicted contrasts made by BrianSurfGCN. The Mann-Whitney U Test

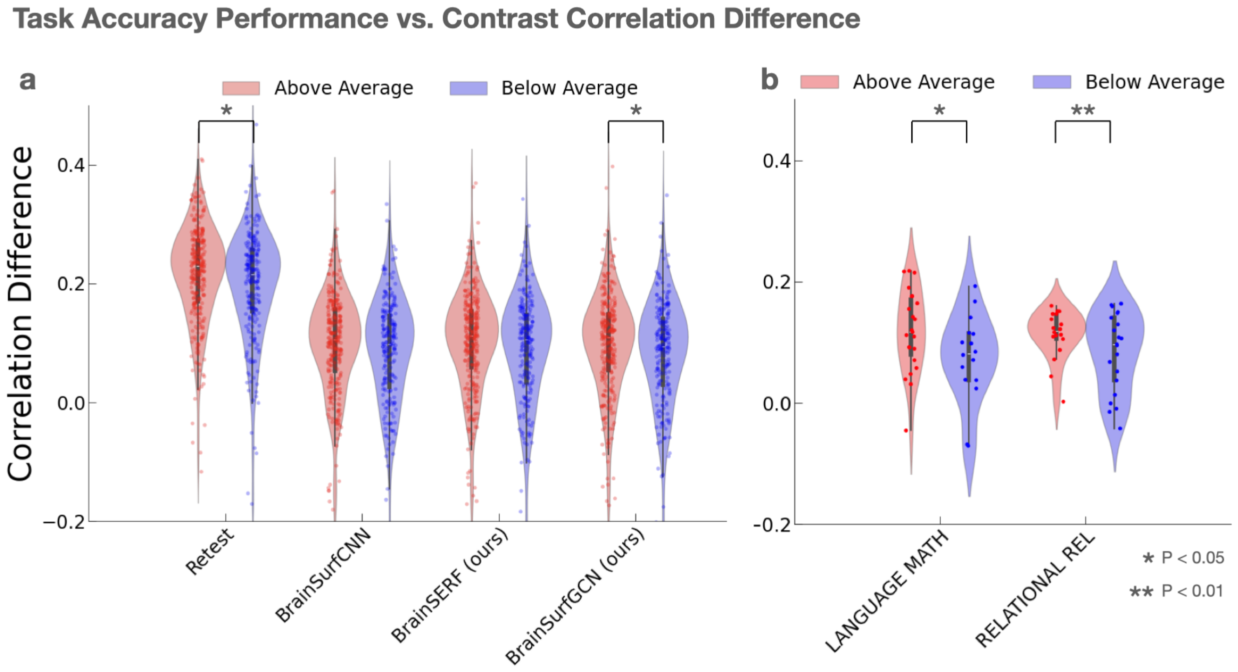

Figure 6: **Correlation difference for above-average and below-average task accuracy groups.** We computed the correlation difference with the ground truth test set for the retest contrasts and each set of model predictions. Selecting only the values associated with tasks with a measured accuracy metric for task performance. We split these sets into above-average and below-average performers. **(a)** For the four sets of contrasts, we found that the retest set and the predictions made by BrainSurfGCN indicate a statistical significance between the two groups in a Mann-Whitney U test. **(b)** The statistical difference in BrainSurfGCN prediction performance was driven by two specific tasks: 'Language: Math' and 'Relational: Relational.'

found p $= 0.018$, U $= 40855.0$ for the retest contrasts, and p $= 0.035$, U $= 40390.0$ for BrainSurfGCN's predicted contrasts. We explored BrainSurfGCN on the single-task level and found two tasks that indicated statistically significant differences between above-average and below-average performers. In these two tasks, we found that the correlation difference was higher in the population that achieved above-average task performance. This indicates that we could predict task activation with BrainSurfGCN more effectively for the above-average performers. For the statistically significant tasks, we found "Language: Math" had p $= 0.035$, U $= 258.0$, and "Relational: Relational" had p $= 0.0042$, U $= 291.0$ in uncorrected comparison tests. Within the retest contrasts, we found that "Relational: Relational" was the only statistically significant task with p $= 0.038$, U $= 263.0$ in uncorrected comparison tests. Being an exploratory procedure, these are analyses centered around hypothesis generation rather than hypothesis testing. Hence, we did not correct for multiple comparisons.

Our second hypothesis considered the effect of noise in the resting-state scan. We posited that more noise in the scan would lead to poorer prediction performance. To test this, we analyze the resting-state temporal signal-to-noise ratio (tSNR) present in the test set of subjects. First, we compute an average tSNR across all test set subjects for each vertex on the mesh. We show this average tSNR plotted on the cortex in Figure 7a. Lower tSNR indicated in blue means that these areas of the mesh had lower tSNR on average for subjects in the test set.

Next, we compute the Dice AUC for two cortical maps: the average tSNR and the average MSE between the ground truth contrasts, and the retest contrasts. This is done across the same threshold interval of top 5% to 50% activation. When thresholding tSNR, we thresholded for the smallest values on the mesh. This allows us to compare regions with the highest prediction error against those with the lowest tSNR. Of the Dice AUC scores we computed, we found the top 5 contrasts with the most overlap to be: 'Language Math vs. Story', 'Working Memory Face vs. Average', 'Motor Tongue vs. Average', 'Gambling Punish vs. Reward', 'Working Memory Body vs. Average'. These contrasts all exhibited a Dice AUC above 0.13, suggesting delta contrast maps have the most overlap between error-prone regions and resting-state tSNR. Figure 7 shows the overlap at a 25% threshold between the average test set MSE and tSNR. We know delta contrast maps show more variation in activation patterns, so we see more widespread activation in Figure 8. This variation makes it very difficult for models to make an accurate prediction since outputs of the models in Figures 5 and 8 suggest a general pattern of smooth prediction.

To further verify that issues in prediction may be associated with noisy properties of delta contrast maps, we computed cortical frequency plots based on thresholded activation. For each subject, we computed the vertices within the top 10% of activated vertices. These vertices were mapped to the cortex and summed up across all test subjects. We then divided this map by the number of test subjects to achieve frequency in activated regions. This frequency plot was compared for all five sets of contrasts: the original test set, the retest contrasts, BrainSurfCNN predictions, BrainSERF predictions, and BrainSurfGCN predictions. We computed the spatial correlation of these frequency plots for each of the four contrasts not belonging to the original test set. These correlation values suggest a marked deficit in correlation associated with delta contrast maps. For example, we examine the Relational task and its three contrasts: Relational Relational, Relational Match, and Relational Match-Relational.

Figure 8 shows a difference between the contrasts taken against baseline and the delta contrast maps. Subject-wise activation is much more variable along the cortex for the delta contrast maps than the other two contrasts. Additionally, the smoother nature of the model predictions suggests that the models do not adeptly predict contrasts with such variability. While the predictions show a slightly more distributed

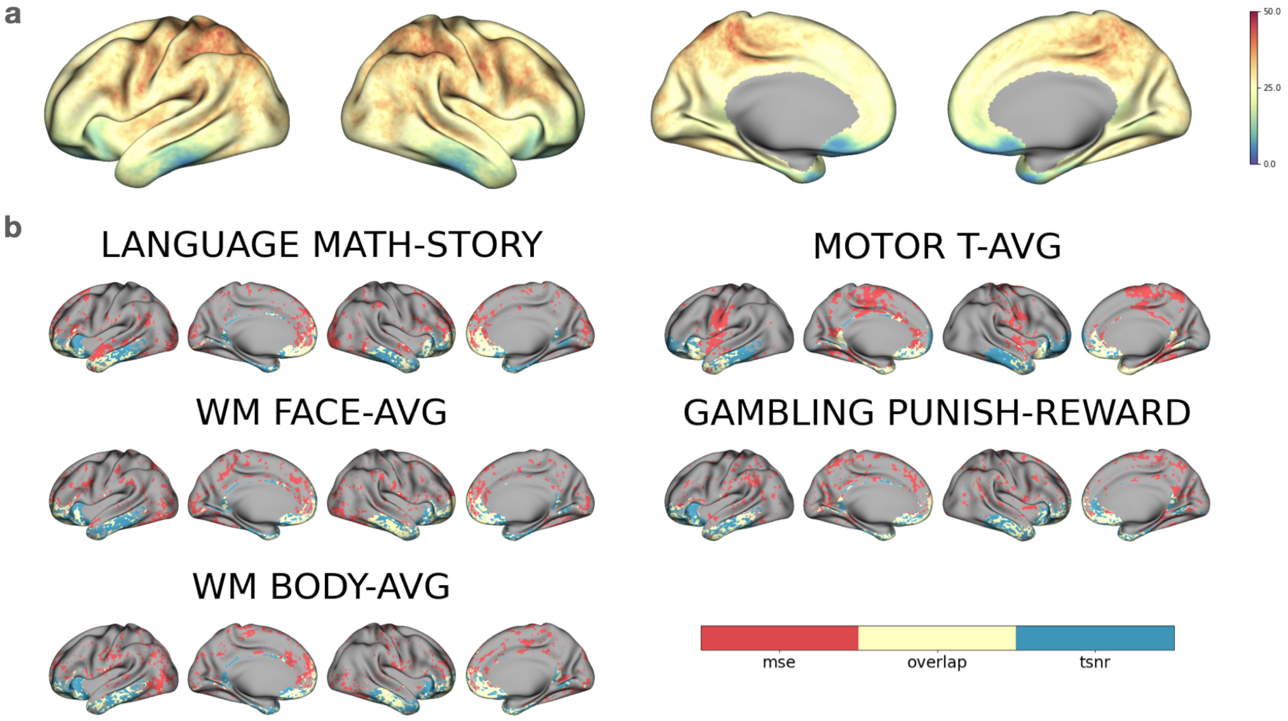

Figure 7: **Relationship between lower tSNR and higher MSE of BrainSurfGCN's predictions on the test set.** **(a)** We computed the average tSNR for each vertex on the mesh across the test set and plotted it on the cortex. **(b)** The top 5 contrasts with the most overlap between the MSE and tSNR using the Dice AUC metric. We measured the overlap between the regions of the highest MSE and lowest tSNR. In this depiction, these maps are shown with a 25% threshold applied to both, allowing us to analyze the specific regions associated with the highest model error.

activation frequency along the cortex, it is much more focused and smooth than the real data. Across the first two relational task maps, shown as an example, the retest spatial correlation was measured at 0.975 and 0.974 for Relational and Matching, respectively. The retest spatial correlation dropped dramatically to 0.626, suggesting that even the retest contrasts may not be able to adequately represent the cortical activation frequency plots for delta contrast maps. The full table of these spatial correlations is available in the Supplementary Tables 7 and 8, and a drop in spatial correlation is consistent for the delta contrast maps.

Additionally, Supplementary Tables 7 and 8 suggest that the noise ceiling aspect of the retest is very dominant when we look at the spatial correlation of frequency plots. To obtain a frequency plot, we threshold each of the contrast maps at 10% and mark a vertex as activated by adding a 1 to the vertex if the vertex is included in the threshold. We then divide the sum of the total activations by the total number of subjects to attain a likelihood of activation for the task. We then do this for the ground truth as well. The Retest contrasts dominated in this measurement, indicating that, from a group level, the

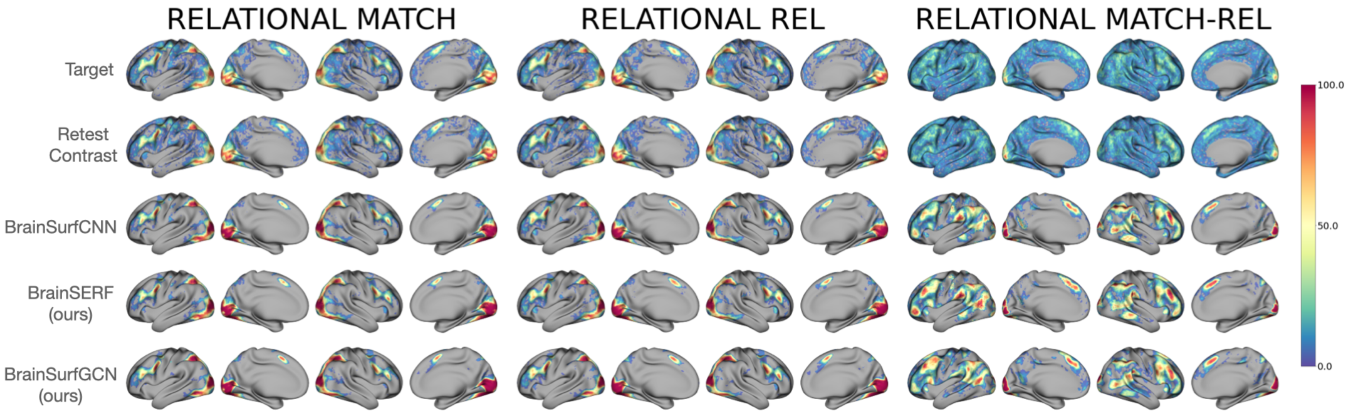

Figure 8: **Cortical frequency plots associated with relational task contrasts.** For each task contrast, we threshold at 10% and count the number of subjects that show activation on the vertex. This gives us a vertex value between 0 and 100, indicating the percentage of subjects with this vertex in the top 10% of activated vertices. Delta contrast maps indicating variability across the population in the activation maps show much more spatial variation compared to both the empirical data and the model-predicted contrasts

Retest maps are still a theoretical upper bound for performance.

4 Discussion

Our models suggest sustained performance and a boost in computational efficiency in individual task contrast prediction through our proposed metrics of spatial correlation, subject identification accuracy, and Dice scores. The enhancements present in BrainSERF are attributed to the inclusion of Squeeze-and-Excitation attention in the mesh convolution layers. This form of attention promotes channel-wise scaling of the channels early on in the model; this can be intuitively understood as an optimized mixing of the IC maps in unison with the U-Net architectures. The improvements in BrainSurfGCN can be attributed to the generalized weight matrices in each layer of the GCN. These weight matrices handle all the feature spaces similarly, so each vertex has the same manipulation as all the others. This allows for a model to generalize feature extraction across the entire mesh.

In our comparative analysis, BrainSERF and BrainSurfGCN demonstrated subtle yet noteworthy enhancements in subject identification accuracy when juxtaposed with the baseline model seen in Table 1. Notably, BrainSurfGCN emerged as a significant leap forward in computational efficiency, reducing the memory demands of BrainSurfCNN's parameters from 12.1MB down to 1.1MB. This reduction marks a significant achievement in making the model more compact and facilitates accelerated training phases on

GPUs with lesser memory capacities. Consequently, the training duration was significantly reduced from approximately 26 to just 6 hours, enhancing the model's usability in resource-constrained environments. Our findings indicate that the performances of BrainSERF and BrainSurfGCN, in terms of our proposed evaluation metrics, are similar to those of the original BrainSurfCNN. This equivalence in performance, coupled with the computational efficiency improvements, underscores the potential of BrainSurfGCN as a more viable option for processing and analyzing fMRI data. Therefore, for applications without computation restrictions, we recommend using BrainSERF, where BrainSurfGCN is more applicable for lightweight computational circumstances.

Our performance debugging is divided into two parts. The first part relates to our hypothesis that a lack of neural engagement could drive a drop in predictability. In our experiments, we found that there may be a significant relationship between task performance (both accuracy and reaction time) and task predictability, at least in a subset of tasks. This relationship between poor prediction performance and poor task performance has also been shown by Gonzalez-Castillo et al., 2015, where the prediction of cognitive states from functional connectivity was impaired by a possible loss of concentration or awareness. We present preliminary evidence for our hypothesis; rigorous future work is needed to better account for task performance during prediction.

Regarding our second hypothesis for the delta contrast map performance deficits, exploratory analysis of the predictions from our models reveals previously undiscussed insights into delta contrast maps and their retest reliability. The discrepancies in performance between a delta contrast map and a contrast calibrated to baseline led us to question why models performed so poorly, especially on specific delta contrast maps. The variability in the delta contrast maps suggests that noise may be present in the activation patterns, so we computed the MSE between the test and retest contrasts. With the vertex-wise error, we analyzed the overlap between tSNR and retest error in the plots shown in Figure 7. We discovered that delta contrast maps had the highest Dice AUC between the error and tSNR compared to contrast maps taken against the baseline. As mentioned, we found the top 5 contrasts with the most overlap: 'Language Math vs. Story', 'Working Memory Face vs. Average', 'Motor Tongue vs. Average', 'Gambling Punish vs. Reward', and 'Working Memory Body vs. Average'. This, coupled with widespread activation seen exclusively in the ground truth and retest maps in Figure 8 suggests that these delta contrast maps are more noise-prone and, thereby, harder to predict.

Our proposed models have several clear limitations. The first is that the additions and changes present in BrainSERF are incremental, and entirely new architectures may show improved performance. This is the issue that led us to explore geometric deep learning. However, the architecture of BrainSurfGCN is straightforward, and we believe newer graph-based learning approaches can yield better performance.

Additionally, we did not explore the generalizability of BrainSurfGCN to graphs of different structures, and it would be interesting to examine how the model performs with a mesh of various sizes and shapes. For example, the original authors of the BrainSurfCNN (Ngo et al., 2022) utilize meshes of various resolutions in the architecture, and these coarser meshes may yield similar performance with a graph network architecture while reducing overall computation. The last limitation lies in how we perform supervised learning. The models are learning to predict all 47 task contrasts, but these models cannot extend to predict unseen task contrasts unless done through fine-tuning on an additional dataset.

Within our limitations, there is plenty of opportunity for future work on the model architecture and explainability of these models. Predicting unseen tasks has been suggested by some of the original authors of BrainSurfCNN (Nguyen et al., 2023). Further, the clinical viability of these predicted task contrasts remains to be explored. While much work has been done to link tfMRI to behavior, more research must be done to determine if and how well model-derived task activation maps link to behavior.

5 Conclusion

Here, we introduced two new models that maintain the state-of-the-art while increasing computational efficiency in subject identification and activation overlap measured with Dice AUC. We recommend BrainSERF for those with high computational capability, whereas BrainSurfGCN is recommended for those with more general-purpose resources. When we examined the performance of these models, we discovered that both the models and the retest contrasts perform more poorly on delta contrast maps, and the analysis of retest error versus tSNR suggests that there is a relationship between retest error and tSNR that is particularly prominent in delta contrast maps. This suggests that data from delta contrast maps are not as reliable as other forms of contrast maps. Additionally, a dive into the performance of BrainSurfGCN revealed that neural engagement is an essential factor in determining the quality of data. Lastly, an ideal clinical application of these methods would mean predicting data for unseen task contrasts. Still, more work remains to be done on connecting synthesized task contrasts to individual behavioral traits.

Author Contribution

Soren Madsen: Conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing of original draft, reviewing and editing, and visualization. Lucina Q. Uddin: Conceptualization, reviewing and editing, and funding acquisition. Jeanette A. Mumford: Conceptualization, reviewing and editing, and funding acquisition. Deanna M. Barch: Conceptualization, reviewing and editing, and funding acquisition. Damien A. Fair: Conceptualization, reviewing and editing, and funding acquisition. Ian H. Gotlib: Conceptualization, reviewing and editing, and funding acquisition. Russell A. Poldrack: Conceptualization, reviewing and editing, supervision, and funding acquisition. Amy Kuceyeski: Conceptualization, validation, reviewing and editing, supervision, and funding acquisition. Manish Saggar: Conceptualization, methodology, investigation, resources, reviewing and editing, supervision, and funding acquisition.

Declaration of Competing Interests

No competing interests are present among the authors of this work.

Data and Code Availability

All code used for this article is made publicly available on GitHub at https://github.com/braindynamicslab/dl-task-contrast-prediction

All data used in this article is publicly available through the Human Connectome Project.

Acknowledgments

This work was supported by an NIH R01 MH127608 and an MCHRI Faculty Scholar Award to Manish Saggar. Data were provided by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

References

Barch, D. M., Burgess, G. C., Harms, M. P., Petersen, S. E., Schlaggar, B. L., Corbetta, M., Glasser, M. F., Curtiss, S., Dixit, S., Feldt, C., et al. (2013). Function in the human connectome: Task-fmri and individual differences in behavior. *Neuroimage*, *80*, 169–189.

Cole, M. W., Ito, T., Bassett, D. S., & Schultz, D. H. (2016). Activity flow over resting-state networks shapes cognitive task activations. *Nature neuroscience*, *19*(12), 1718–1726.

Deary, I. J., Simonotto, E., Meyer, M., Marshall, A., Marshall, I., Goddard, N., & Wardlaw, J. M. (2004). The functional anatomy of inspection time: An event-related fmri study. *Neuroimage*, *22*(4), 1466–1479.

Emch, M., Von Bastian, C. C., & Koch, K. (2019). Neural correlates of verbal working memory: An fmri meta-analysis. *Frontiers in human neuroscience*, *13*, 180.

Fey, M., & Lenssen, J. E. (2019). Fast graph representation learning with PyTorch Geometric. *ICLR Workshop on Representation Learning on Graphs and Manifolds*.

Finn, E. S. (2021). Is it time to put rest to rest? *Trends in cognitive sciences*, *25*(12), 1021–1032.

Ghojogh, B., & Crowley, M. (2019). The theory behind overfitting, cross validation, regularization, bagging, and boosting: Tutorial. *arXiv preprint arXiv:1905.12787*.

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., et al. (2013). The minimal preprocessing pipelines for the human connectome project. *Neuroimage*, *80*, 105–124.

Gonzalez-Castillo, J., Hoy, C. W., Handwerker, D. A., Robinson, M. E., Buchanan, L. C., Saad, Z. S., & Bandettini, P. A. (2015). Tracking ongoing cognition in individuals using brief, whole-brain functional connectivity patterns. *Proceedings of the National Academy of Sciences*, *112*(28), 8762–8767. https://doi.org/10.1073/pnas.1501242112

Hearne, L. J., Mill, R. D., Keane, B. P., Repovš, G., Anticevic, A., & Cole, M. W. (2021a). Activity flow underlying abnormalities in brain activations and cognition in schizophrenia. *Science Advances*, *7*(29), eabf2513. https://doi.org/10.1126/sciadv.abf2513

Hearne, L. J., Mill, R. D., Keane, B. P., Repovš, G., Anticevic, A., & Cole, M. W. (2021b). Activity flow underlying abnormalities in brain activations and cognition in schizophrenia. *Science Advances*, *7*(29), eabf2513. https://doi.org/10.1126/sciadv.abf2513

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *International conference on machine learning*, 448–456.

Jiang, C., Huang, J., Kashinath, K., Marcus, P., Niessner, M., et al. (2019). Spherical cnns on unstructured grids. *arXiv preprint arXiv:1901.02039*.

Jiang, R., Zuo, N., Ford, J. M., Qi, S., Zhi, D., Zhuo, C., Xu, Y., Fu, Z., Bustillo, J., Turner, J. A., Calhoun, V. D., & Sui, J. (2020). Task-induced brain connectivity promotes the detection of individual differences in brain-behavior relationships. *NeuroImage*, *207*, 116370. https://doi.org/https://doi.org/10.1016/j.neuroimage.2019.116370

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Liu, R., Lehman, J., Molino, P., Petroski Such, F., Frank, E., Sergeev, A., & Yosinski, J. (2018). An intriguing failing of convolutional neural networks and the coordconv solution. *Advances in neural information processing systems*, *31*.

Milletari, F., Navab, N., & Ahmadi, S.-A. (2016). V-net: Fully convolutional neural networks for volumetric medical image segmentation. *2016 fourth international conference on 3D vision (3DV)*, 565–571.

Ngo, G. H., Khosla, M., Jamison, K., Kuceyeski, A., & Sabuncu, M. R. (2022). Predicting individual task contrasts from resting-state functional connectivity using a surface-based convolutional network. *NeuroImage*, *248*, 118849. https://doi.org/https://doi.org/10.1016/j.neuroimage.2021.118849

Nguyen, M., Ngo, G. H., & Sabuncu, M. R. (2023). Zero-shot learning of individualized task contrast prediction from resting-state functional connectomes. *arXiv preprint arXiv:2310.14105*.

Pagliaccio, D., Middleton, R., Hezel, D., Steinman, S., Snorrason, I., Gershkovich, M., Campeas, R., Pinto, A., Van Meter, P., Simpson, H. B., et al. (2019). Task-based fmri predicts response and remission to exposure therapy in obsessive-compulsive disorder. *Proceedings of the National Academy of Sciences*, *116*(41), 20346–20353.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, *32*.

Quah, S., Jo, B., Geniesse, C., Uddin, L., Mumford, J., Barch, D., Fair, D., Gotlib, I., Poldrack, R., & Saggar, M. (2024). A data-driven latent variable approach to validating the research domain criteria framework. *bioRxiv*. https://doi.org/10.1101/2024.01.31.577486

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241.

Smith, S. M., Beckmann, C. F., Andersson, J., Auerbach, E. J., Bijsterbosch, J., Douaud, G., Duff, E., Feinberg, D. A., Griffanti, L., Harms, M. P., et al. (2013). Resting-state fmri in the human connectome project. *Neuroimage*, *80*, 144–168.

Smith, S. M., Fox, P. T., Miller, K. L., Glahn, D. C., Fox, P. M., Mackay, C. E., Filippini, N., Watkins, K. E., Toro, R., Laird, A. R., et al. (2009). Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the national academy of sciences*, *106*(31), 13040–13045.

Tavor, I., Jones, O. P., Mars, R. B., Smith, S., Behrens, T., & Jbabdi, S. (2016). Task-free mri predicts individual differences in brain activity during task performance. *Science*, *352*(6282), 216–220.

Van Essen, D. C., Glasser, M. F., Dierker, D. L., Harwell, J., & Coalson, T. (2012). Parcellations and hemispheric asymmetries of human cerebral cortex analyzed on surface-based atlases. *Cerebral cortex*, *22*(10), 2241–2262.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., & Ugurbil, K. (2013). The wu-minn human connectome project: An overview [Mapping the Connectome]. *NeuroImage*, *80*, 62–79. https://doi.org/https://doi.org/10.1016/j.neuroimage.2013.05.041

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wang, J., Lv, P., Wang, H., & Shi, C. (2021). Sar-u-net: Squeeze-and-excitation block and atrous spatial pyramid pooling based residual u-net for automatic liver segmentation in computed tomography. *Computer Methods and Programs in Biomedicine*, *208*, 106268.

Ward, N., Brown, M., Thompson, A., & Frackowiak, R. (2003). Neural correlates of outcome after stroke: A cross-sectional fmri study. *Brain*, *126*(6), 1430–1448.

Yeo, B. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., Roffman, J. L., Smoller, J. W., Zöllei, L., Polimeni, J. R., et al. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology*.

Zhang, Y., Tetrel, L., Thirion, B., & Bellec, P. (2021). Functional annotation of human cognitive states using deep graph convolution. *NeuroImage*, *231*, 117847.

A Supplementary Information

A.1 Additional Models Explored

A.1.1 BaggedSurfCNN and BaggedSurfGCN

In the development of BaggedSurfCNN, we adopt the core architectural framework of BrainSurfCNN Ngo et al., 2022, ensuring consistency in model structure to facilitate a direct comparison of performance metrics. The innovation in our approach lies in the application of a bagging ensemble methodology, coupled with a 5-fold cross-validation training strategy. This combination is designed to enhance the model's generalization capability and mitigate overfitting, a common challenge in deep learning models applied to neuroimaging data. For a comparison to our new architectures, we repeated the same approach using BrainSurfGCN as well.

The 5-fold cross-validation process divides the available dataset into five distinct subsets. In each fold of the cross-validation, four subsets are utilized for training, and the remaining subset is reserved for validation. This procedure is repeated five times, with each subset serving as the validation set once, ensuring that every data point contributes to both training and validation phases. Through this method, we generate five fully trained BaggedSurfCNN models, each having been exposed to slightly different training and validation data splits. This diversity in training conditions is key to the ensemble approach, as it encourages the development of models with varied predictive strengths and biases Ghojogh and Crowley, 2019.

Upon completion of the training phase, we employ a bagging strategy for ensemble prediction. This involves aggregating the predictions from all five models for a given input to form a single ensemble prediction. Specifically, we compute the average of the predicted task contrasts across all models. This averaging process is predicated on the assumption that while individual models may make errors, the ensemble, through averaging, is likely to converge towards a more accurate and robust prediction. This is because errors made by individual models, assuming they are uncorrelated, are likely to cancel out when averaged, thereby reducing the overall prediction error. The ensemble prediction for the task contrast, thus, represents a consensus across the five models, mitigating the impact of any single model's potential biases or inaccuracies. In both of these bagged approaches, we noticed no significant improvements over the existing training strategy.

A.1.2 BrainSurfATN

The introduction of the Transformer architecture Vaswani et al., 2017 has revolutionized the field of sequence modeling, especially within the domain of natural language processing (NLP). A cornerstone of this architecture is the multi-head attention mechanism, which facilitates the modeling of complex dependencies between sequence elements, such as words in a sentence. This mechanism enables the model to focus on different parts of the sequence simultaneously, thereby capturing nuanced relationships between tokens that are not necessarily adjacent to each other.

In the context of our research, we adapt the multi-head attention framework to the analysis of brain surface meshes. Recognizing the potential of this mechanism to enhance the representation of spatial data, we introduce a multi-head attention layer into the Residual Pooling Block of our model. This adaptation is predicated on the conceptualization of the mesh's vertices as a sequence, akin to tokens in NLP. Each vertex represents a unique point on the brain's surface, and the relationships between these vertices are crucial for understanding the underlying anatomical and functional organization of the brain.

By implementing a multi-head attention layer, our goal was to contextualize each vertex within the global structure of the mesh. This means that for any given vertex, the model can assess its significance and relationships in the context of the entire mesh, rather than considering it in isolation or solely in relation to its immediate neighbors. This approach allows for a more nuanced representation of the brain's surface, capturing both local and distant anatomical features. It is particularly advantageous for tasks that require a comprehensive understanding of the brain's geometry and functional connectivity, as it enables the model to dynamically adjust the importance it assigns to each vertex based on the broader context of the mesh sequence.

A.2 Evaluation of Additional Models

Table 4: Subject identification test accuracy across varying ICs, MSE Only

ICs	15	25	50	100
BrainSurfCNN	75.5	75.9	75.7	76.2
BrainSERF	76.5	**78.0**	77.6	N/A
BrainSurfGCN	**77.9**	77.9	**78.9**	**79.7**
BrainSurfATN	57.9	55.2	42.0	37.2
BaggedSurfCNN	N/A	61.9	N/A	N/A
BaggedSurfGCN	N/A	75.3	N/A	N/A

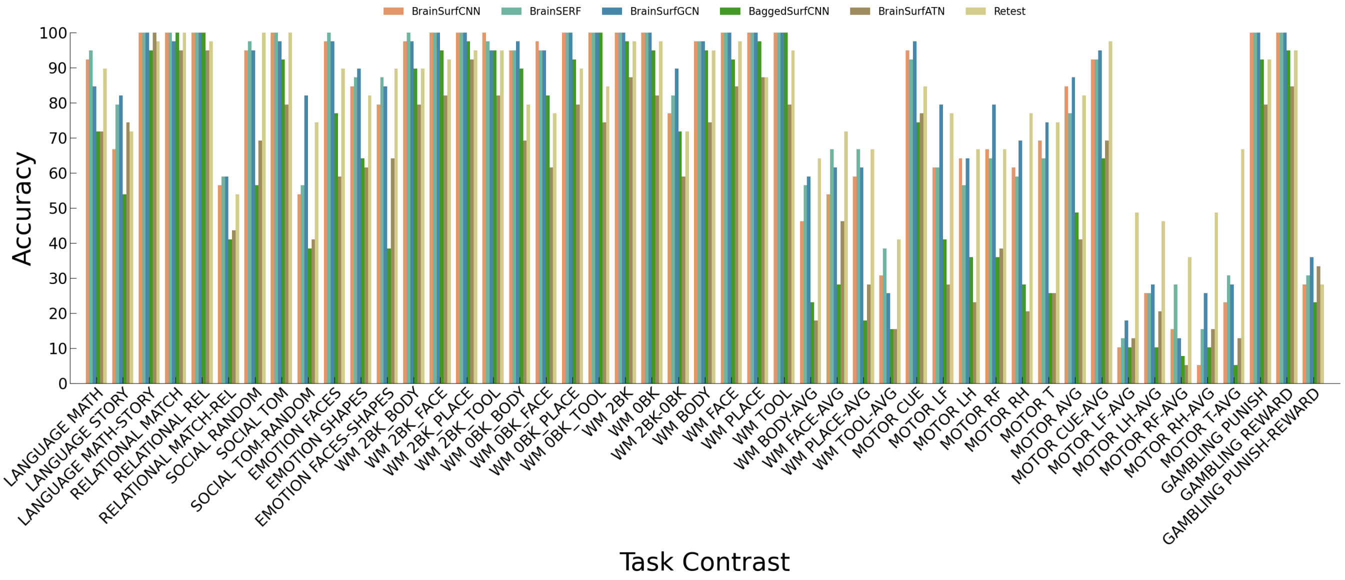

Figure 9: Best model accuracy on all task contrasts, MSE Training Only

A.3 Additional Results of All Final Models

We have computed the Dice AUC for all tasks contrasts for the retest contrasts and the predicted contrasts from all of the final models.

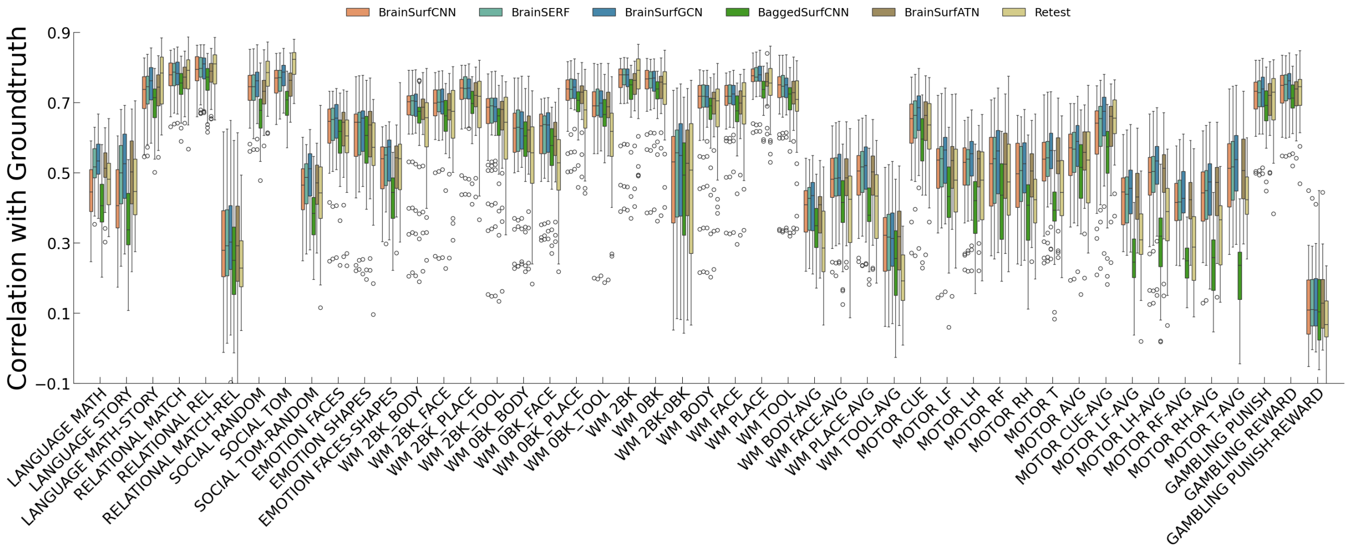

Figure 10: Best model correlation across all task contrasts, MSE Training Only

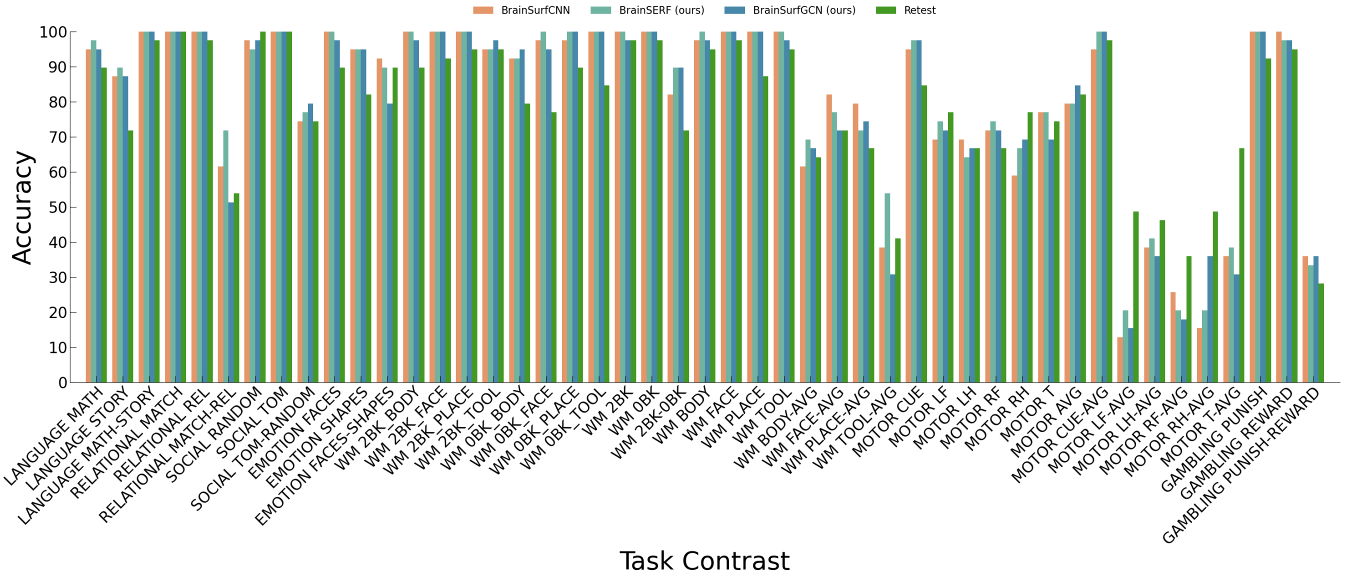

Figure 11: Subject identification accuracy across all task contrasts.

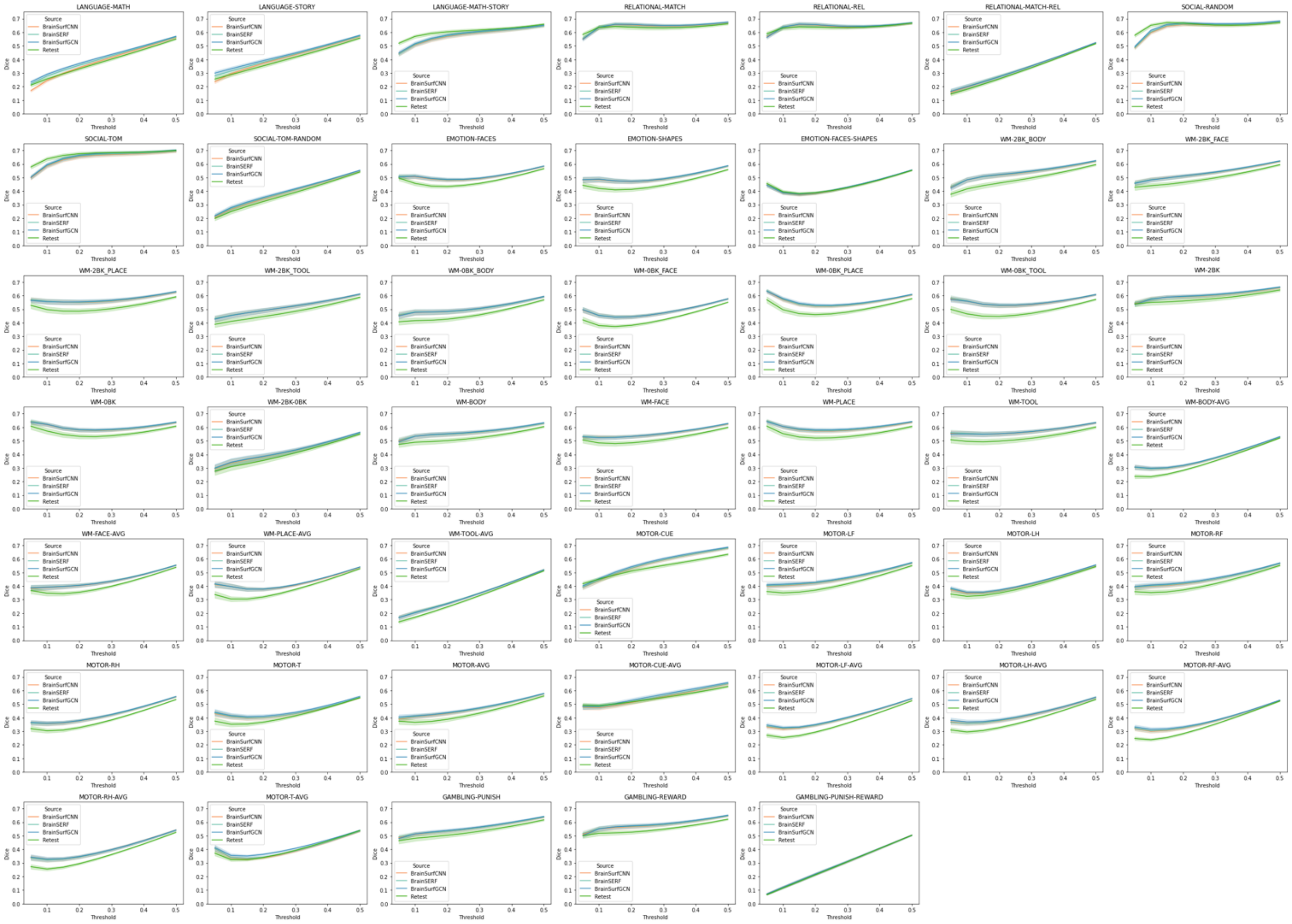

Figure 12: Dice scores for all final models across all task contrasts.

Table 5: Dice AUC Across All Language, Relational, Social, Emotional, and Working Memory Tasks

Contrast	Retest	BrainSurfCNN	BrainSERF	BrainSurfGCN
LANGUAGE MATH	**0.182**	0.144	0.163	0.157
LANGUAGE STORY	**0.176**	0.153	0.162	0.167
LANGUAGE MATH-STORY	0.247	0.274	0.276	**0.280**
RELATIONAL MATCH	0.227	**0.267**	0.265	0.264
RELATIONAL REL	**0.219**	**0.219**	0.217	0.218
RELATIONAL MATCH-REL	**0.138**	0.115	0.118	0.105
SOCIAL RANDOM	0.259	0.258	0.259	**0.260**
SOCIAL TOM	0.271	0.298	0.300	**0.305**
SOCIAL TOM-RANDOM	0.125	0.144	**0.145**	0.142
EMOTION FACES	0.201	0.216	0.215	**0.220**
EMOTION SHAPES	0.195	0.19	0.192	**0.197**
EMOTION FACES-SHAPES	0.201	0.225	**0.229**	0.226
WM 2BK_BODY	0.17	**0.212**	0.211	0.209
WM 2BK_FACE	**0.222**	0.202	0.2	0.201
WM 2BK_PLACE	0.153	**0.27**	**0.27**	0.267
WM 2BK_TOOL	0.18	0.191	0.191	**0.192**
WM 0BK_BODY	0.243	0.269	0.269	**0.273**
WM 0BK_FACE	0.187	0.22	**0.221**	0.22
WM 0BK_PLACE	0.144	**0.258**	0.257	0.252
WM 0BK_TOOL	0.228	0.238	0.236	**0.239**
WM 2BK	0.249	**0.261**	**0.261**	0.259
WM 0BK	0.249	**0.294**	0.293	**0.294**
WM 2BK-0BK	**0.142**	0.137	0.139	0.141
WM BODY	0.249	**0.277**	0.276	0.276
WM FACE	0.218	**0.234**	0.233	0.233
WM PLACE	0.141	**0.282**	**0.282**	0.279
WM TOOL	**0.227**	0.222	0.22	0.221
WM BODY-AVG	**0.151**	0.141	0.14	0.149
WM FACE-AVG	0.126	0.192	**0.199**	0.193
WM PLACE-AVG	0.105	0.196	**0.199**	0.19
WM TOOL-AVG	0.125	**0.137**	0.136	0.132

Table 6: Dice AUC Across All Motor and Gambling Tasks

Contrast	Retest	BrainSurfCNN	BrainSERF	BrainSurfGNN
MOTOR CUE	0.207	0.213	**0.217**	**0.217**
MOTOR LF	0.154	**0.21**	0.209	0.208
MOTOR LH	0.152	**0.155**	0.154	0.149
MOTOR RF	0.136	0.191	0.193	**0.195**
MOTOR RH	0.135	**0.145**	0.143	**0.145**
MOTOR T	0.155	0.205	**0.209**	0.202
MOTOR AVG	0.15	0.219	**0.22**	0.218
MOTOR CUE-AVG	**0.197**	0.165	0.167	0.172
MOTOR LF-AVG	0.144	0.153	0.154	**0.155**
MOTOR LH-AVG	0.135	**0.172**	**0.172**	0.167
MOTOR RF-AVG	0.125	0.135	**0.137**	**0.137**
MOTOR RH-AVG	0.126	0.146	0.144	**0.149**
MOTOR T-AVG	0.151	0.157	0.157	**0.158**
GAMBLING PUNISH	**0.272**	0.251	0.249	0.252
GAMBLING REWARD	**0.246**	0.206	0.205	0.209
GAMBLING PUNISH-REWARD	**0.129**	0.098	0.097	0.104

Table 7: Spatial correlation of percentage of subjects showing top 10% activation between predicted and ground truth contrasts associated language, relational, social, emotional, and working memory tasks.

Contrast	Retest	BrainSurfCNN	BrainSERF	BrainSurfGCN
LANGUAGE MATH	**0.772**	0.581	0.701	0.717
LANGUAGE STORY	**0.843**	0.691	0.757	0.783
LANGUAGE MATH-STORY	**0.954**	0.908	0.907	0.914
RELATIONAL MATCH	**0.975**	0.958	0.959	0.955
RELATIONAL REL	**0.974**	0.963	0.965	0.958
RELATIONAL MATCH-REL	**0.626**	0.554	0.558	0.591
SOCIAL RANDOM	**0.977**	0.939	0.941	0.938
SOCIAL TOM	**0.973**	0.932	0.932	0.933
SOCIAL TOM-RANDOM	**0.769**	0.724	0.73	0.751
EMOTION FACES	**0.943**	0.902	0.908	0.901
EMOTION SHAPES	**0.934**	0.901	0.902	0.897
EMOTION FACES-SHAPES	**0.915**	0.796	0.807	0.792
WM 2BK_BODY	**0.922**	0.918	0.919	0.91
WM 2BK_FACE	**0.931**	0.912	0.911	0.905
WM 2BK_PLACE	**0.948**	0.937	0.939	0.934
WM 2BK_TOOL	**0.919**	0.895	0.896	0.888
WM 0BK_BODY	**0.93**	0.905	0.905	0.9
WM 0BK_FACE	**0.916**	0.88	0.881	0.869
WM 0BK_PLACE	**0.949**	0.928	0.931	0.926
WM 0BK_TOOL	**0.946**	0.931	0.932	0.924
WM 2BK	**0.954**	0.945	0.945	0.939
WM 0BK	**0.962**	0.949	0.95	0.944
WM 2BK-0BK	**0.833**	0.815	0.82	0.821
WM BODY	**0.946**	0.931	0.93	0.924
WM FACE	**0.939**	0.923	0.923	0.915
WM PLACE	**0.956**	0.946	0.948	0.944
WM TOOL	**0.95**	0.933	0.934	0.926
WM BODY-AVG	**0.792**	0.72	0.722	0.718
WM FACE-AVG	**0.905**	0.815	0.823	0.813
WM PLACE-AVG	**0.885**	0.839	0.841	0.829
WM TOOL-AVG	**0.639**	0.554	0.569	0.577

Table 8: Spatial correlation of percentage of subjects showing top 10% activation between predicted and ground truth contrast sets associated with motor and gambling tasks.

Contrast	Retest	BrainSurfCNN	BrainSERF	BrainSurfGNN
MOTOR CUE	**0.93**	0.858	0.864	0.869
MOTOR LF	**0.887**	0.839	0.845	0.846
MOTOR LH	**0.885**	0.757	0.764	0.75
MOTOR RF	**0.891**	0.827	0.832	0.839
MOTOR RH	**0.87**	0.789	0.795	0.788
MOTOR T	**0.903**	0.815	0.814	0.826
MOTOR AVG	**0.886**	0.851	0.855	0.851
MOTOR CUE-AVG	**0.938**	0.887	0.892	0.897
MOTOR LF-AVG	**0.851**	0.707	0.717	0.712
MOTOR LH-AVG	**0.891**	0.768	0.762	0.776
MOTOR RF-AVG	**0.845**	0.673	0.686	0.69
MOTOR RH-AVG	**0.857**	0.725	0.723	0.72
MOTOR T-AVG	**0.893**	0.664	0.672	0.701
GAMBLING PUNISH	**0.932**	0.919	0.922	0.911
GAMBLING REWARD	**0.945**	0.931	0.936	0.92
GAMBLING PUNISH-REWARD	**0.391**	0.182	0.179	0.19