Human Brain Mapping

WILEY

RESEARCH ARTICLE OPEN ACCESS

Efficient Deep Learning Models for Predicting Individualized Task Activation From Resting-State Functional Connectivity

Soren J. Madsen[1] | Young-Eun Lee[1] | Shaun K. L. Quah[1] | Lucina Q. Uddin[2] | Jeanette A. Mumford[3] | Deanna M. Barch[4] | Damien A. Fair[5] | Ian H. Gotlib[3] | Russell A. Poldrack[3] | Amy Kuceyeski[6] | Manish Saggar[1]

[1]Department of Psychiatry, Stanford University, Stanford, California, USA | [2]Department of Psychiatry, University of California, Los Angeles, California, USA | [3]Department of Psychology, Stanford University, Stanford, California, USA | [4]Department of Psychology, Washington University in St. Louis, St. Louis, Missouri, USA | [5]Department of Pediatrics, University of Minnesota, Minneapolis, Minnesota, USA | [6]Department of Radiology, Weill Cornell Medicine, New York, New York, USA

Correspondence: Manish Saggar (saggar@stanford.edu)

Received: 3 February 2026 | **Revised:** 19 April 2026 | **Accepted:** 15 May 2026

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

ABSTRACT

Deep learning models have demonstrated the potential to predict task-evoked brain activation from resting-state functional magnetic resonance imaging, offering a pathway toward individualized brain mapping without requiring task-based data. In this study, we systematically evaluate architectural strategies for improving the efficiency and scalability of such models. Using data from the Human Connectome Project, we replicate the BrainSurfCNN framework and introduce two extensions: BrainSERF, which incorporates channel-wise attention through squeeze-and-excitation modules, and BrainSurfGCN, a graph-based model that leverages cortical mesh topology for efficient message passing. Across multiple evaluation metrics, including spatial correlation, Dice score, Dice AUC, and subject identification accuracy, all models achieve comparable predictive performance. Despite similar accuracy, the proposed models offer distinct advantages. BrainSERF provides modest improvements in capturing individual-specific features, while BrainSurfGCN achieves substantial reductions in model size and training time, highlighting a favorable trade-off between performance and computational efficiency. Beyond architectural comparisons, we investigate factors driving variability in prediction accuracy. We find that behavioral task performance, resting-state data quality, and inter-subject variability in task activation jointly constrain prediction fidelity. In particular, contrasts with lower signal reliability and higher variability exhibit reduced predictability across all models. Together, these findings demonstrate that incorporating topological and functional structural priors can improve the efficiency of deep learning models without sacrificing accuracy, while also emphasizing that prediction performance is fundamentally limited by the reliability of the underlying neural signals.

1 | Introduction

One of the foundational questions in cognitive neuroscience is whether the brain's intrinsic organization, observable during rest, encodes information predictive of how that same brain will respond during active cognitive tasks. If intrinsic functional connectivity reflects stable traits that constrain task-evoked activation, then models trained on resting-state data should be able to accurately predict task contrast maps at the individual level (Cole et al. 2016; Ngo et al. 2022; Osher et al. 2016; Tavor et al. 2016). This hypothesis has significant implications for understanding brain-behavior relationships,

Soren J. Madsen and Young-Eun Lee contributed equally to this work.

characterizing individual differences, and developing precision neuroimaging tools, particularly for populations that are unable to perform scanner-based tasks. At the same time, validating such models provides a data-driven approach to test assumptions about the relationship between intrinsic and task-evoked activity.

Prior work has demonstrated that the spatial organization of resting-state networks closely mirrors the brain's task-evoked architecture (Smith et al. 2009; Smith-Collins et al. 2015; Sui et al. 2009), suggesting that intrinsic functional connectivity may serve as a latent scaffold for cognitive function. This raises the possibility that resting-state functional magnetic resonance imaging (rsfMRI) contains sufficient information to reconstruct task activation maps at the individual level, effectively using the brain at rest to infer how it behaves in action. In addition to its practical utility in cases where task data are unavailable or infeasible (e.g., clinical or pediatric populations), such predictions offer a powerful test of the stability, specificity, and behavioral relevance of intrinsic connectivity (Chow et al. 2025; Hearne et al. 2021; Jiang et al. 2020; Savage et al. 2024; Serin et al. 2025). Recent reviews have further summarized advances in this area and highlighted its expanding methodological scope and applications (Bernstein-Eliav and Tavor 2024), extending connectome-based prediction to include mechanistic frameworks and deep learning approaches. The feasibility of this approach has also been demonstrated in clinical populations, including pre-surgical mapping and psychiatric cohorts (Jones et al. 2017; Tik et al. 2021), highlighting its potential translational utility.

This line of inquiry also has significant implications for translational research frameworks such as the NIMH Research Domain Criteria (RDoC), which seek to redefine mental illness in terms of dimensional, neurobiologically grounded constructs (Cuthbert and Insel 2013; Insel 2014). Task-fMRI has traditionally played a central role in identifying brain circuits linked to RDoC domains, yet no single individual completes all relevant tasks (Quah, Madsen, et al. 2025; Quah, Jo, et al. 2025). Accurate rest-to-task prediction could enable virtual approximation of these activation patterns, potentially allowing comprehensive, individualized neural profiling without requiring data acquisition across all RDoC task domains. Thus, predictive models offer both a mechanistic window into brain organization and a scalable tool for next-generation clinical neuroscience.

Foundational work by Tavor et al. (2016) and Cole et al. (2016) demonstrated that task activation maps could be predicted from resting-state functional connectivity using linear models, establishing a baseline for rest-to-task inference. These studies showed that intrinsic connectivity patterns contain behaviorally relevant information. Building on this foundation, Ngo et al. 2022 introduced BrainSurfCNN, a deep learning model that integrates U-Net-style architecture with surface-based mesh convolutions to significantly improve individual-level prediction accuracy. While technically impressive, such models raise two central questions that remain unresolved: (1) Can architectural changes enhance the fidelity or efficiency of these predictions? And (2) why are some individuals or task contrasts more predictable than others?

To address the first question, we examine whether these architectural changes can enhance the fidelity or efficiency of rest-to-task predictions. We introduce two new models built upon distinct principles. The first, BrainSERF, augments BrainSurfCNN with squeeze-and-excitation (SE) modules (Hu et al. 2018), which implement dynamic channel-wise feature reweighting. This mechanism performs adaptive channel-wise feature reweighting that modulates information flow based on contextual salience. The second model, BrainSurfGCN, utilizes graph convolution (Kipf and Welling 2016) and models information propagation over the cortical mesh using graph-based operations by treating the cortical surface as a graph and explicitly modeling localized, neighborhood-based information flow, thereby explicitly modeling localized, neighborhood-based information flow on the cortical mesh.

To address the second question, we investigate why rest-to-task prediction succeeds for some individuals or task contrasts but not others. Specifically, we examine whether variability in model performance may reflect meaningful differences in cognitive engagement, neural signal reliability, and the degree to which intrinsic connectivity scaffolds task-evoked activity. To address this, we evaluate how behavioral performance and scan quality influence prediction accuracy, and we examine the cortical distribution of prediction errors in relation to known anatomical noise profiles. These analyses aim to reveal when and where brain dynamics and individual traits facilitate or constrain successful decoding of task-related activity from resting-state data.

2 | Method

2.1 | Dataset and Preprocessing

We used de-identified, publicly available data from the HCP dataset (Van Essen et al. 2013) to train our models in the same way as BrainSurfCNN. The study was approved by the Stanford University Institutional Review Board.

Specifically, as input to our models, we used the FIX-cleaned, 3 T rsfMRI data acquired in four 14.4-min runs, each with 1200 time points per session per subject. 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), where rsfMRI preprocessing includes highpass filtering, MELODIC ICA, and FIX for component classification. 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 4800 time points split into eight samples of 600 time points from which the functional connectivity is determined. These eight samples allow us to extend the training set of our data by uniformly sampling these eight chunks in each training epoch.

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. We created new training datasets derived from 15, 25, 50, and 100 independent components

using these spatial components. As 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. 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.

Additionally, we only included subjects with all four resting-state runs and all 47 task contrasts, which gave us 919 subjects for the training set, 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.

Following the filtering approach described by Ngo et al. (2022), our test–retest sample consisted of 39 out of 46 available subjects who met the data completeness criteria, specifically having all four rsfMRI runs and 47 tfMRI contrasts. We refer to the contrasts from the first visit as the "test set" and those from the second visit as the "repeat dataset," adopting BrainSurfCNN terminology. The test set participants do not overlap with those in our training set. We used the repeat dataset as an optimal predictor of ground-truth task activation (i.e., non-model-based or model-free). A contrast taken from the same subject in a different session represents the best possible model for individual variability, providing an approximate noise ceiling, particularly at the group level, although individual-level comparisons may deviate due to measurement variability. Notably, resting-state data from the repeat sessions were not used as model input, but were instead used as a reference to estimate the noise ceiling. In addition, we included a group-average baseline as a non-individualized reference. Specifically, for each task contrast, we computed the average activation map across training subjects and used this as a prediction for each individual subject. This baseline reflects population-level activation patterns while removing subject-specific variability.

In the proposed deep learning framework, the input data are structured as multi-channel fsLR polyhedral meshes, each consisting of 32,492 vertices, to accommodate the high-dimensional spatial topology of cerebral cortex representations. Here, "fsLR" denotes FreeSurfer-derived cortical surface meshes aligned to a standardized template (fsLR template) across the left and right hemispheres. 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

We evaluated three models, BrainSurfCNN, BrainSERF, and BrainSurfGCN, for predicting task contrast maps. These models differ in how they represent and propagate spatial information on the cortical surface. BrainSurfCNN models local spatial structure via mesh convolutions. BrainSERF extends this framework by enhancing feature selectivity through channel-wise recalibration. BrainSurfGCN, in contrast, propagates information across the cortical graph using message passing. We describe each model in detail below.

2.2.1 | BrainSurfCNN

BrainSurfCNN is a mesh-based encoder–decoder convolutional network that maps resting-state IC representations to task contrast maps using hierarchical surface convolutions. 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 Jiang et al. (2019).

The architecture adopts a hierarchical encoder-decoder structure composed of downsampling and upsampling pathways, which leverage mesh convolutional operations to effectively capture spatial features across multiple resolutions. Importantly, it incorporates a Residual Pooling Block (implemented as ResPoolBlock), which combines residual connections, mesh convolutions, and pooling operations to enhance robust multi-scale feature learning. 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 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.

2.2.2 | BrainSERF

BrainSERF extends BrainSurfCNN by incorporating channel-wise feature recalibration through squeeze-and-excitation (SE) modules. The SE attention mechanism has emerged as a pivotal innovation in the field of computer vision, significantly enhancing the representational power of convolutional neural networks 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 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 applied 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). The network additionally incorporates the Residual Pooling Block, as described in BrainSurfCNN, to further enhance hierarchical feature learning through residual connections and mesh-based pooling.

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. The BrainSERF architecture is described in Figure 1a, and additional training hyperparameters, command-line arguments, and example usage scripts are provided on our GitHub repository: https://github.com/braindynamicslab/dl-task-contrast-prediction.

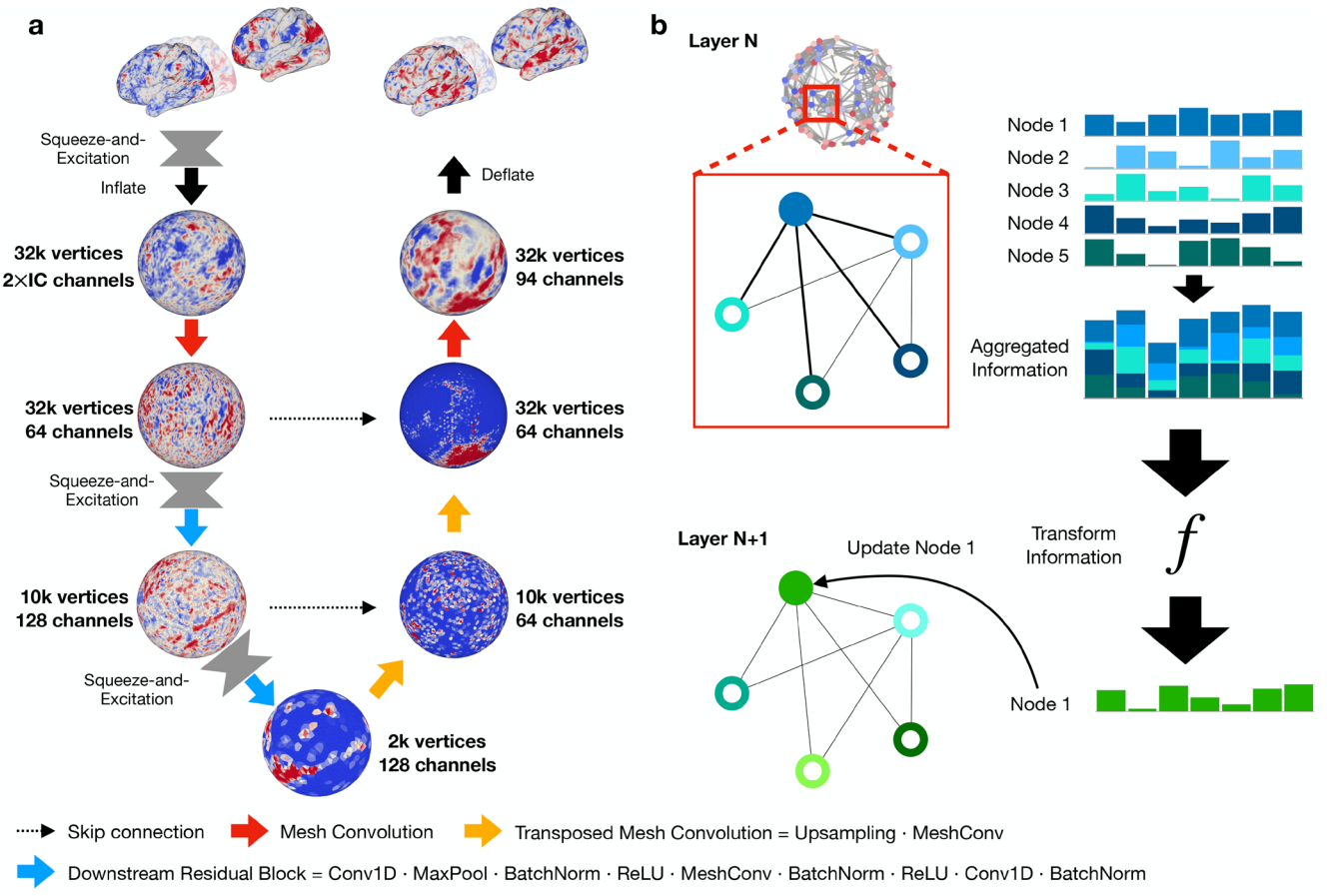

FIGURE 1 | Overview of BrainSERF and BrainSurfGCN architecture for cortical mesh-based fMRI analysis. (a) In the BrainSERF model, input surface-based fMRI data are processed through a hierarchical mesh convolutional network, incorporating SE modules and skip connections. Vertices and channel dimensions are progressively transformed using mesh convolution, residual blocks, and upsampling operations. (b) In each BrainSurfGCN layer, node features in the cortical mesh are aggregated, transformed, and updated according to graph connectivity. Aggregated node information is passed through nonlinear transformations to generate updated node representations for subsequent layers. This framework enables multi-scale integration and hierarchical feature extraction from full-brain functional mesh data.

2.2.3 | BrainSurfGCN

BrainSurfGCN is a graph convolutional network that operates directly on the cortical mesh, using message passing to propagate information across neighboring vertices.

In predicting task contrast activation maps from rsfMRI, employing graph neural networks offers several compelling advantages that align with the intrinsic properties of brain data and the objectives of neuroimaging analysis (Zhang et al. 2021). 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 graph-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 $\mathbb{R}^{|\mathcal{V}|\times C}$, 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). The 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 a BDLayer (described in Figure S1). The model consists of multiple graph-based layers, each composed of 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:

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

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

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

where $\hat{A} = A + I$ denotes an adjacency matrix with added self-loops, $\hat{D}$ denotes the diagonal degree matrix of $\hat{A}$, 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 three 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 $\mathbb{R}^{32{,}492\times(2*50+3)}$. Figure 1b shows our architecture from a high-level perspective, complete with the input and output feature spaces.

2.3 | 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 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.

After performing some preliminary sanity checks (e.g., correlation between prediction and ground truth), 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

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

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

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.

2.4 | Evaluation

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

The spatial 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 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 (Dice score) 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 score 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 obtain a threshold-independent summary, we further compute the Dice AUC by integrating Dice scores across all evaluated thresholds. This provides a more robust assessment of spatial overlap that does not depend on a single threshold choice while still capturing both the spatial distribution and relative magnitude of 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. Moreover, we computed this accuracy for each quantity of ICs used in training.

3 | Results

In this section, we address two core questions central to our investigation: (1) Can architectural changes improve the fidelity and efficiency of predicting task-evoked brain activation from resting-state connectivity? and (2) Why does prediction performance vary across individuals and cognitive tasks?

To this end, we organize our results in a three-part sequence: (i) Architectural Design and Trade-offs: We first evaluate predictive performance across three model architectures, including BrainSurfCNN (baseline), BrainSERF, and BrainSurfGCN, highlighting efficiency improvements. Model predictions are compared against both a repeat scan from the same subject, serving as a noise ceiling and a group-average baseline, serving as a non-individualized reference. (ii) Model Evaluation: We then assess subject-level prediction specificity through a subject identification analysis, offering a complementary measure of how well each model captures individualized activation patterns. (iii) Mechanisms of Variability: Finally, we examine factors that drive individual and contrast-level differences in prediction performance, including task engagement, data quality, and inter-subject variability in activation patterns. All results were computed using 50 ICs for consistency.

3.1 | Architectural Design and Trade-Offs

We first evaluated how well each model, including BrainSurfCNN, BrainSERF, and BrainSurfGCN, could predict individual-level task activation maps from resting-state connectivity patterns. As shown in Figure 2 for one representative participant, all three models qualitatively captured key spatial features of task-evoked responses, such as somatomotor activation during the Motor task and activation in association cortices during the Social task. BrainSurfGCN achieved comparable spatial localization and alignment with ground truth, while substantially reducing model complexity and computational cost. The thresholded task activation maps for the other IC settings (e.g., 15, 25, and 100) across all predictions for the seven task contrasts are presented in Figures S2, S3, and S4. Performance was not substantially affected by the choice of IC dimensionality.

To quantitatively evaluate model performance, we computed Dice score between predicted and ground-truth activation maps across multiple thresholds (5%–50%, in 5% increments), following the procedure established in Ngo et al. (2022). Thresholding allows us to isolate task-relevant signal while mitigating the influence of noise. Figure 3 shows example outputs for three thresholds (10%, 25%, and 50%) from the Social Cognition: Theory of Mind contrast, illustrating how prediction sensitivity changes with threshold level.

We computed Dice scores across thresholds from 5% to 50% in 5% steps to characterize how prediction quality changes with varying activation selectivity. As shown in Figure 4, these curves provide a clearer view of model behavior across activation intensities. Across different task contrasts, we observe three representative patterns: (1) convergence between repeat and group average performance, (2) comparable performance between model predictions and repeat, and (3) improved performance of model predictions over both repeat and group average. In several contrasts, model predictions slightly exceeded the repeat dataset, which represents the empirical noise ceiling. This pattern, also noted by Ngo et al. 2022, likely arises because repeat scans contain session-to-session variability and small spatial shifts in activation, lowering their measurable overlap. The smoother and less noisy nature of model predictions can therefore yield higher Dice scores without violating theoretical limits. All Dice scores for every final model across all task contrasts are provided in Figure S6.

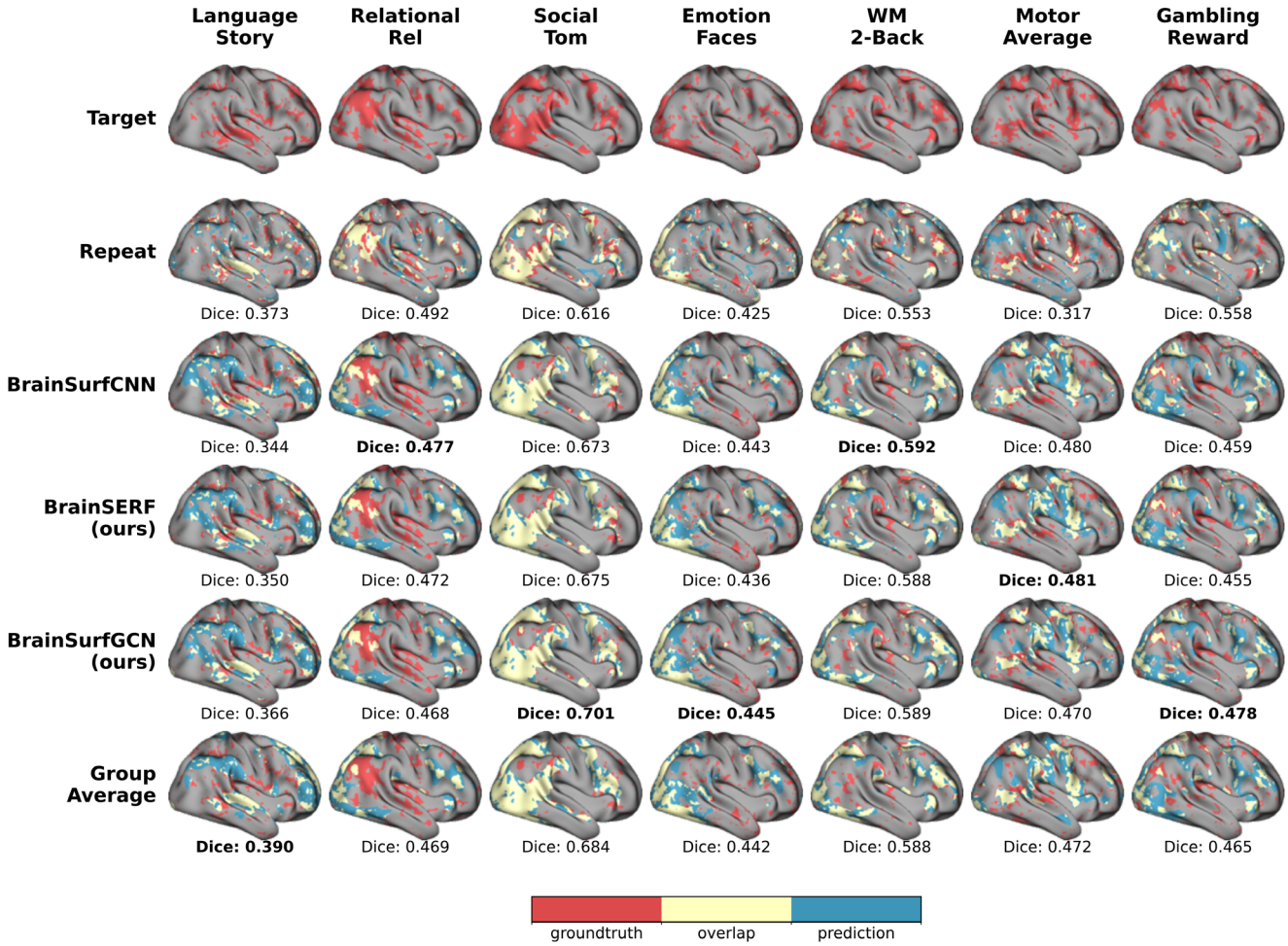

FIGURE 2 | Thresholded task activation maps for seven task contrasts using a threshold of 25%. We compare ground truth, repeat, group average, and models' predicted activation for an individual (subject 917,255). The thresholded activation maps on the right hemisphere (lateral view) represent, respectively, Language: Story, Relational Processing: Relational, Social Cognition: Theory of Mind, Emotion: Faces, Working Memory: 2-Back, Motor: Average, and Gambling: Reward. Ground truth (target), repeat scans, group average, and predictions from BrainSurfCNN, BrainSERF, and BrainSurfGCN are presented with colors indicating ground truth, model prediction, and overlap. Dice scores are displayed beneath each row, quantifying prediction fidelity across thresholds. The highest Dice value within each column is bolded to highlight the best-performing models.

To summarize performance across thresholds, we calculated the area under the Dice curve (Dice AUC), with a maximum possible value of 0.45 based on perfect overlap across the integration range. These values are shown in Figure 5 and provide a concise metric for comparing model performance across tasks and individuals. We also examined unthresholded predictions for baseline reference.

Finally, extended results, including per-task Dice AUC and additional visualizations, are available in Tables S1 and S2 and Figure S7–S15, providing a more granular view of how each model performs across contrasts and individuals.

Finally, we quantified the computational footprint of each model to highlight trade-offs between performance and efficiency. Table 1 reports the number of trainable parameters across architectures and ICs. Notably, BrainSurfGCN achieved substantial reductions in parameters. This is attributable to its mesh-based design, which allows parameter sharing across nodes and avoids the need to store full mesh gradients. Importantly, model size remained largely stable across different IC configurations, since operations scale with fixed hidden-state sizes rather than the number of ICs directly.

3.2 | Evaluation of Predictive Fidelity—Subject-Level Specificity

To evaluate the individual specificity of model predictions, we examined both the spatial correlation between predicted and ground-truth task activation maps and the subject identification accuracy. This analysis assesses whether the models capture unique, individualized features of brain function that distinguish one subject from another.

We first computed the spatial correlation difference: for each predicted contrast, we compared its correlation with the corresponding ground-truth contrast from the same subject versus those from different subjects. A high correlation difference indicates that the model more accurately captures idiosyncratic

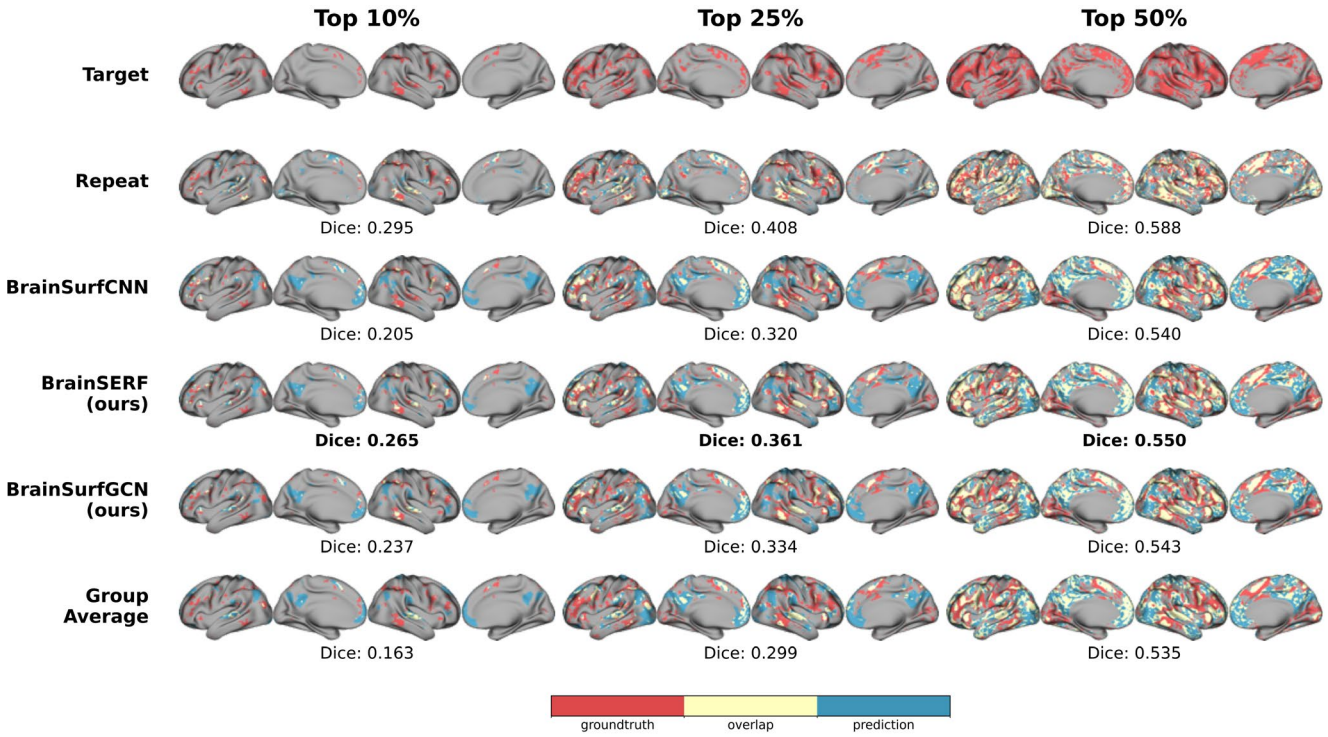

FIGURE 3 | Thresholded task activation maps for Social Cognition: Theory of Mind. Activation maps for an individual (subject 917,255) are thresholded at three levels (top 10%, top 25%, and top 50%) to illustrate how effectively each model captures the spatial distribution of task-evoked responses. Ground truth (target), repeat scans, group average, and predictions from BrainSurfCNN, BrainSERF, and BrainSurfGCN are shown with colors indicating ground truth, model prediction, and overlap. Dice scores are displayed beneath each row, quantifying prediction fidelity across thresholds. The highest Dice value within each column is bolded to highlight the best-performing models.

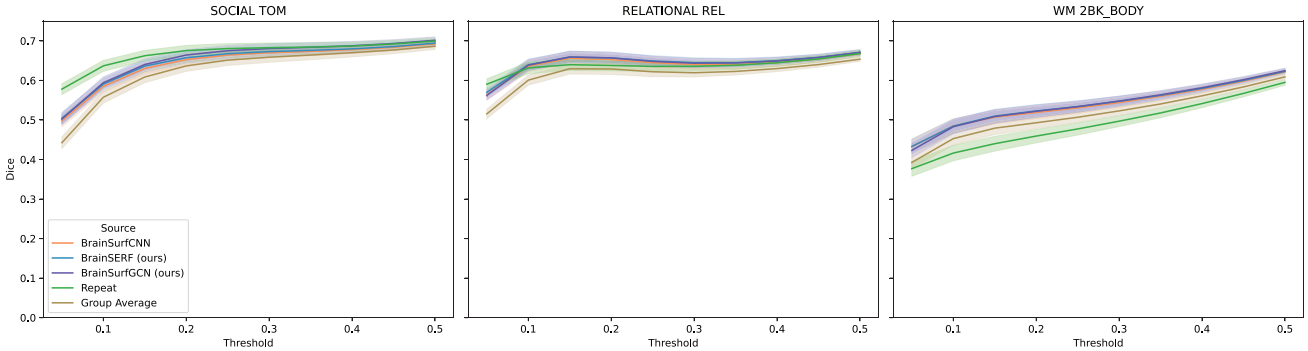

FIGURE 4 | Dice score plots for three selected tasks showing variations in performance across task contrasts. Each plot compares model predictions with the repeat condition and group average baseline across thresholds from 5% to 50% for Social Cognition: Theory of Mind, Relational: Relational, and Working Memory: 2-Back Body. The shaded regions denote standard deviations in the dice scores across test subjects.

features of the target individual. As shown in Figure 6, all three models exhibit strong within-subject correlation relative to others, suggesting that they reliably capture individually localized patterns of task-evoked activity.

Next, we performed a subject identification task: for each predicted contrast, we computed its correlation with all ground-truth contrasts across the test set and evaluated whether the highest correlation matched the true subject. This classification-based metric offers a concrete measure of the identifiability of individuals from predicted brain maps. To examine the effect of input dimensionality, we systematically varied the number of ICA components (15, 25, 50, and 100 ICs). Average subject identification accuracies across task contrasts are summarized in Table 2.

Interestingly, all models achieved comparable accuracy levels but remained generally below the repeat benchmark. The average subject identification accuracy for repeat contrasts was 92.9% across 47 task contrasts, highlighting the strong test–retest reliability of the data. Notably, in a few tasks (e.g., Language: Math and Language: Story), model predictions slightly exceeded identifiability from repeat scans (S5).

The performance did not consistently improve with increasing ICA dimensionality. Instead, subject identification accuracy remained relatively stable across resolutions, with no clear benefit beyond moderate dimensionality. This suggests that increasing input dimensionality alone does not necessarily translate to improved individual-level prediction.

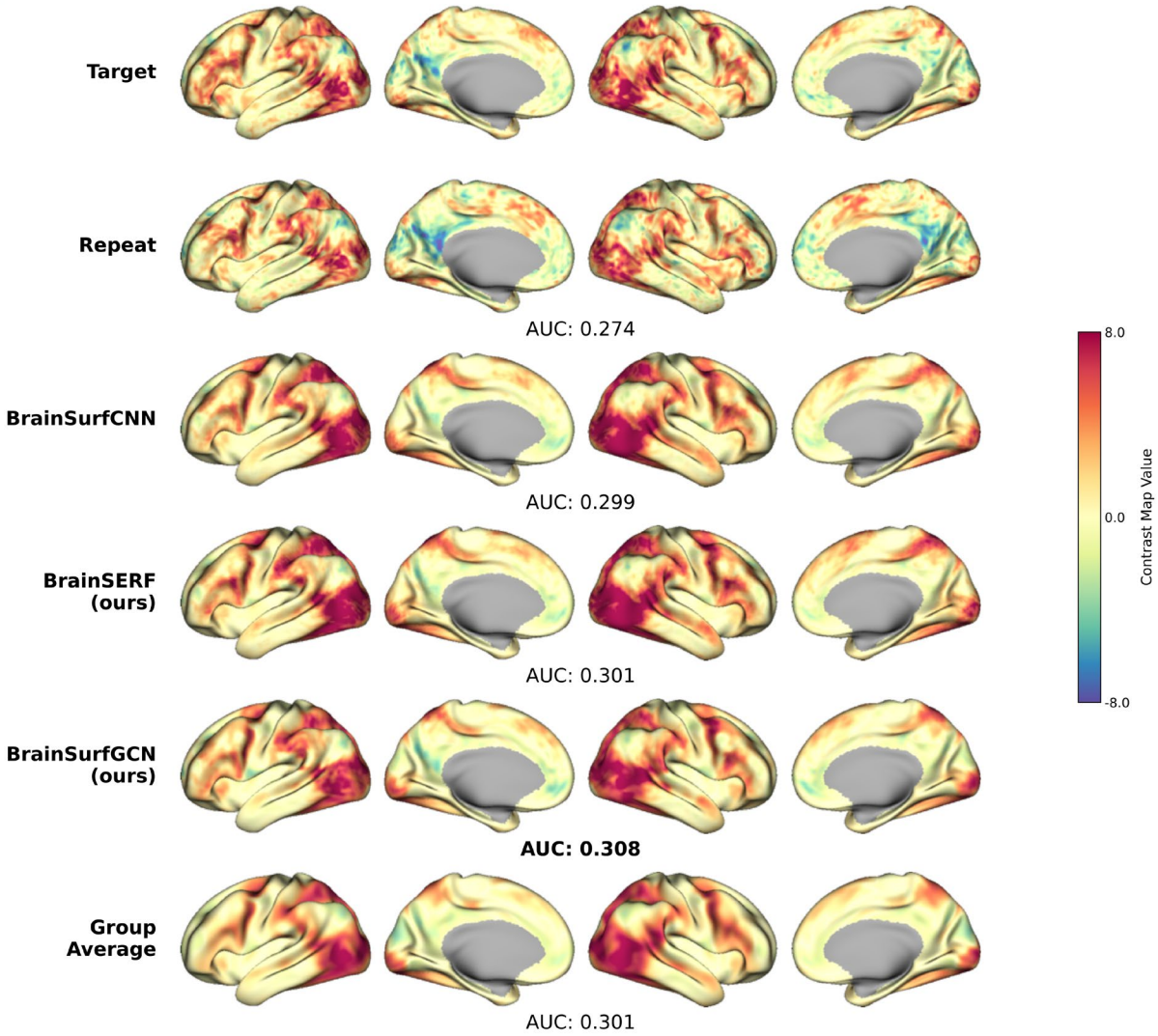

FIGURE 5 | Unthresholded task activation for Social Cognition: Theory of Mind We compare ground truth, repeat, group average, and model-predicted activation for an individual (subject 917,255). Model predictions are much smoother on the whole than the ground truth or repeat activations. Though smoother, the models perform higher in the Dice AUC than the repeat contrasts. All activation maps are displayed in z score units.

TABLE 1 | 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**

Note: Values in bold indicate the most efficient model (lowest number of parameters) for each set of ICs.

3.3 | Mechanisms of Variability in Predictability

Having established that all three architectures can generate individualized task activation patterns, we next asked why prediction accuracy varies across individuals and contrasts. We focused on three potential sources of variability grounded in cognitive neuroscience: task engagement, resting-state data quality, and inter-subject variability in task-evoked activation.

3.3.1 | Task Engagement Influences Predictability

We first examined whether behavioral task engagement modulates the success of rest-to-task prediction. To this end, we focused on the held-out test set ($n = 39$) and stratified participants based on their task performance (e.g., accuracy or behavioral variability). If a participant performs poorly on a task, reflected in lower accuracy, or greater behavioral variability, the neural representations associated with that task may be weaker or less

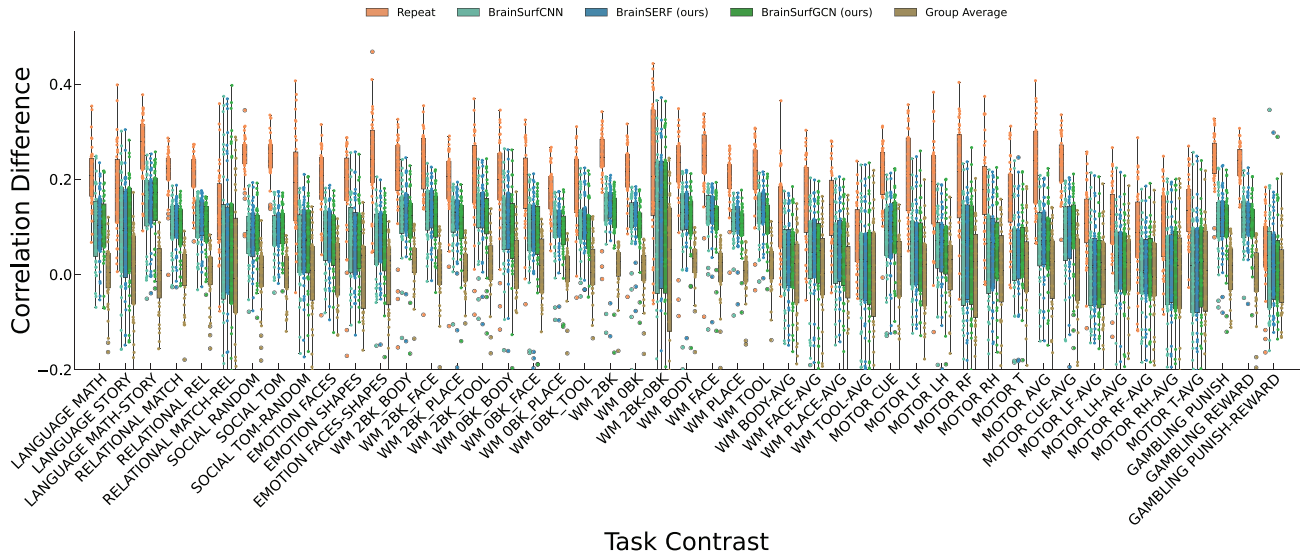

FIGURE 6 | Difference in spatial correlation of predicted contrast to ground truth contrast of the same subject versus the mean spatial correlation of predicted contrast to ground truth contrasts of all other test set subjects. The correlation metric primarily measures the linear relationship between 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. Model predictions are compared against both the repeat condition and the group average baseline. 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 2 | 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	79.8	**82.8**
BrainSurfGCN (ours)	79.5	79.5	**80.7**	**79.9**	82.4
Repeat (Benchmark)			**92.9**		

Note: Values in bold indicate the highest value within each column (across rows, excluding the last row).

consistently expressed, reducing the model's ability to reconstruct individual activation patterns. For the behavioral performance analysis, we restricted our focus to a subset of 14 task contrasts that included valid accuracy metrics, allowing us to stratify subjects into high (above median) and low (below median) performance groups.

Although several contrasts showed nominal significance prior to correction, most effects did not survive multiple-comparison correction using the Benjamini–Hochberg FDR procedure. At the model level (Figure 7a), nominally significant group differences were observed for the repeat data and BrainSurfGCN, with above-average performers showing higher correlation differences than below-average performers; however, these effects did not survive FDR correction, potentially reflecting limited statistical power.

Motivated by these nominal effects, we further examined task-specific patterns using BrainSurfGCN (Figure 7b). At the task level, only two contrasts, Working Memory: 2-Back Body and Working Memory: 2-Back Place, showed borderline significance after correction, with higher correlation differences in the above-average group.

All task contrasts are shown in Figure S16, illustrating the full distribution of effects across tasks. Overall, these results suggest that the relationship between behavioral performance and predictability is limited, task-specific, and modest in magnitude. Accordingly, these analyses should be interpreted as exploratory and hypothesis-generating.

3.3.2 | Resting-State Signal Quality Shapes Prediction Error

We next examined whether resting-state data quality constrains prediction performance. Specifically, we evaluated the relationship between regional temporal signal-to-noise ratio (tSNR) and prediction error, measured as MSE.

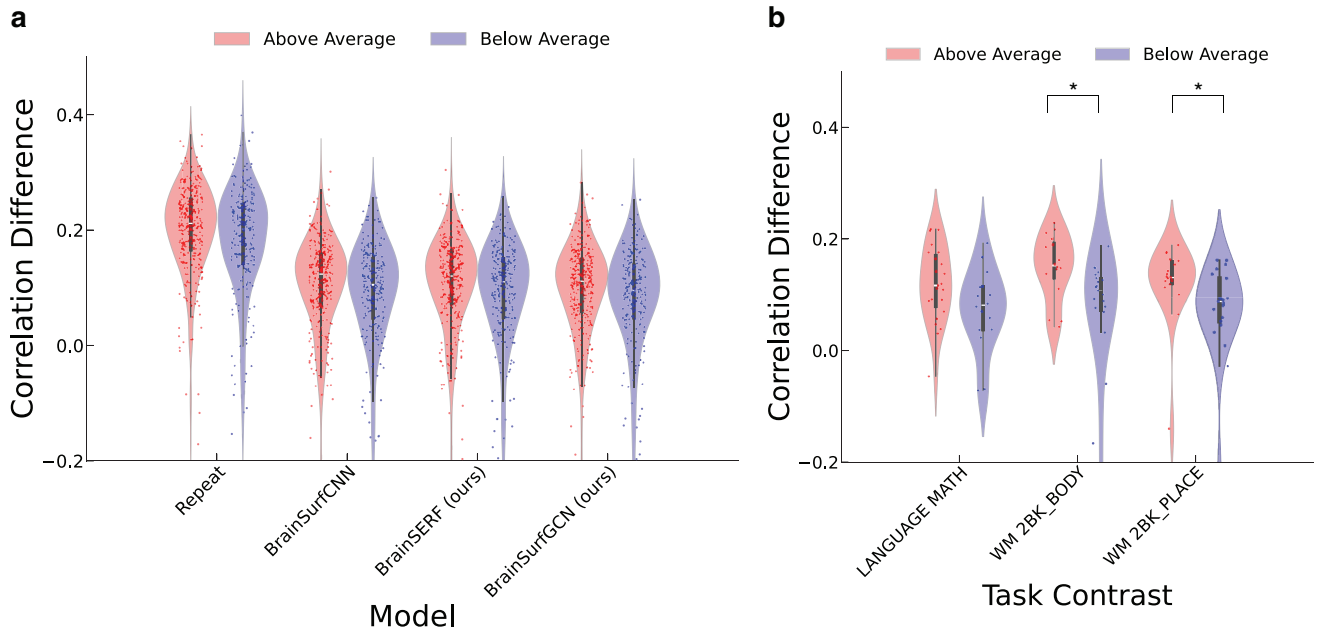

FIGURE 7 | Correlation difference for above- and below-average task accuracy groups. Subjects were divided into above-average (red) and below-average (blue) behavioral performance groups for task contrasts with an available accuracy metric, and correlation differences were compared between groups. (a) Across models (repeat, BrainSurfCNN, BrainSERF, and BrainSurfGCN), no group differences remained significant after multiple-comparison correction (Benjamini–Hochberg FDR). Each point represents a subject's correlation difference for a single task contrast (aggregated across all 14 task contrasts and 39 subjects; $n = 546$ points per model). (b) Three representative task contrasts are shown (BrainSurfGCN). Among all contrasts, only Working Memory: 2-Back Body and Working Memory: 2-Back Place showed significant group differences after FDR correction, with higher correlation differences in the above-average group. Here, each point represents a single subject ($n = 39$ per contrast).

We first computed the average tSNR across subjects for each cortical vertex (Figure 8a). Regions with lower tSNR, such as the orbitofrontal cortex and temporal poles, are known to be susceptible to susceptibility artifacts and physiological noise. We then compared the spatial distribution of low-tSNR regions with areas showing high prediction error, computed by comparing model predictions with repeat contrasts.

Across all 47 contrasts, delta contrasts ($n = 16$) showed significantly higher error-to-tSNR Dice AUC than single-condition contrasts ($n = 31$; mean 0.108 vs. 0.088; one-sided Mann–Whitney U test, $p = 0.004$; 10,000-permutation test, $p = 0.001$; Cliff's $\delta = 0.48$, large effect; Figure S17). All five contrasts with the highest overlap were delta contrasts, including Language: Math—Story, Working Memory: Face—Avg, and Motor: Tongue—Avg. These results indicate that rest-to-task prediction performance is partly limited by the intrinsic noise characteristics of resting-state data, especially for contrasts requiring subtle condition-specific distinctions.

3.3.3 | Variability in Cortical Activation Across Subjects and Contrasts

Finally, we asked whether inter-subject variability in task-evoked activation constrains predictability. Prior studies have shown that contrasts involving subtle cognitive differences, particularly delta contrasts, exhibit less spatial consistency across individuals than single-condition contrasts (Seghier and Price 2016).

To quantify this, we generated activation-frequency maps for each contrast by computing, for each subject, the top 10% of activated vertices and then calculating the proportion of subjects showing activation at each vertex. These maps were computed for the empirical data (test and repeat) and for predictions from all three models.

As illustrated in Figure 9, delta contrasts showed markedly greater spatial variability than single-condition contrasts. Even the repeat contrasts showed low spatial correlation for delta contrasts (e.g., Relational Match—Relational: $r = 0.626$), underscoring their low test–retest reliability. In contrast, single-condition contrasts such as Relational: Relational and Relational: Match showed high repeatability ($r \approx 0.97$), reflecting stable and stereotyped activation patterns.

The models reproduced these broad patterns: predictions were smoother and more spatially constrained than the empirical data and showed reduced correspondence for delta contrasts relative to single-condition contrasts. This suggests that high inter-subject variability, combined with lower reliability of delta contrasts, limits the upper bound of predictability for these contrasts, even for the empirical repeat maps.

Tables S3 and S4 provide full spatial correlation matrices, further highlighting the dominant influence of the repeat-derived noise ceiling on contrast-level predictability.

Together, these analyses reveal that variability in rest-to-task prediction arises from a combination of cognitive factors (task engagement), technical factors (resting-state data quality), and neurobiological factors (inter-subject variability in cortical recruitment). These findings emphasize that prediction performance is not solely determined by model architecture, but is fundamentally

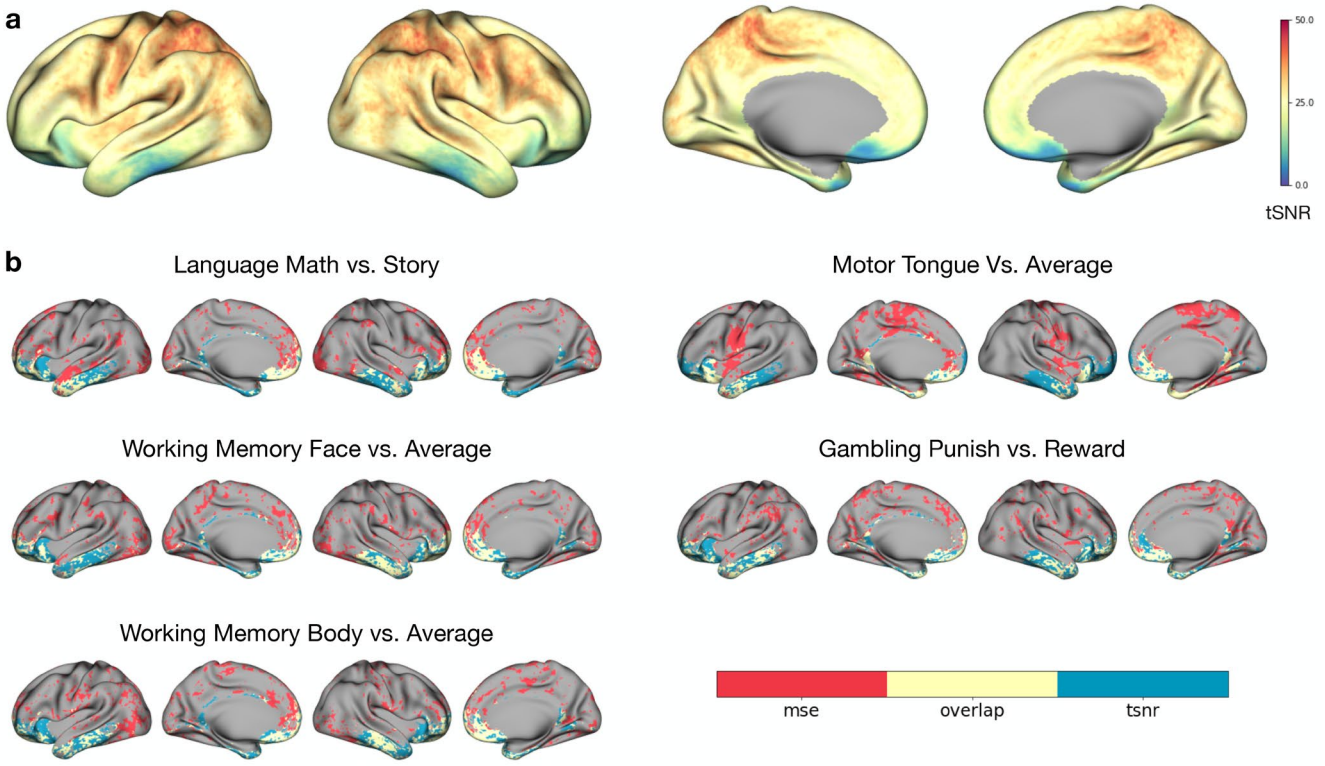

FIGURE 8 | Relationship between lower tSNR and higher MSE of BrainSurfGCN's predictions on the test set. (a) We computed the average tSNR for each vertex on the mesh across the test set and plotted it on the cortex. (b) The top five contrasts with the most overlap between the MSE and tSNR using the Dice 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.

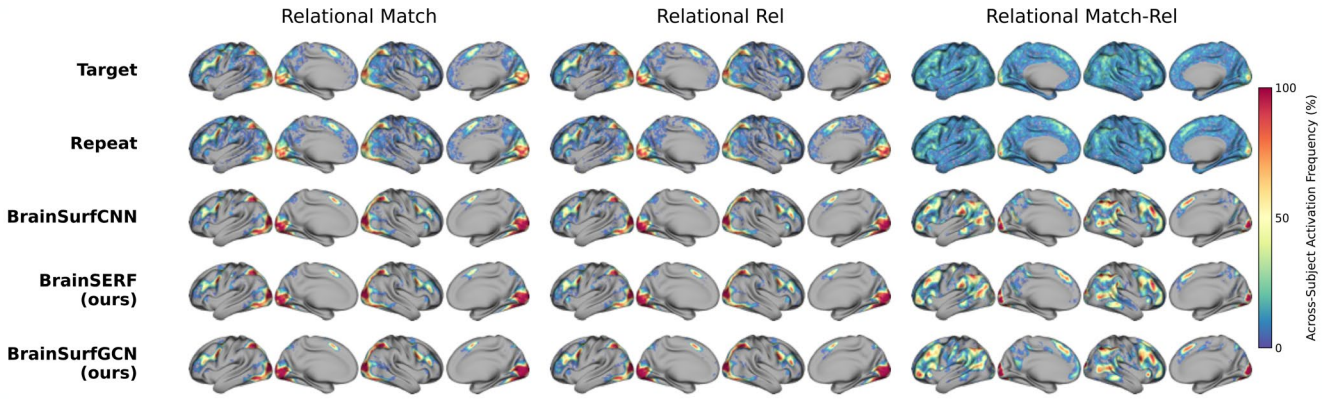

FIGURE 9 | Cortical frequency plots associated with relational task contrasts. For each task contrast, we threshold at 10% and count the number of subjects that show activation on the vertex. This gives us a vertex value between 0 and 100, indicating the percentage of subjects with this vertex in the top 10% of activated vertices. Delta contrast maps indicating variability across the population show substantially greater spatial variation compared to both the repeat and the model-predicted contrasts.

shaped by the stability and reliability of the underlying neural signals.

4 | Discussion

In this work, we reproduced the predictive performance of BrainSurfCNN and introduced two novel variants, including BrainSERF, which integrates a channel-wise attention mechanism for adaptive feature refinement, and BrainSurfGCN, which implements a graph-based representation to explicitly model the organization of brain networks. These models were designed in response to two central questions raised in the introduction: (1) Can deep learning models of resting-state to task prediction be made more architecturally efficient and of higher fidelity? And (2) What are the key drivers of variability in individual-level prediction performance?

4.1 | Model Innovations: Attention and Graph-Based

4.1.1 | Architectures

BrainSERF incorporates a channel-wise attention mechanism (SE block) that allows the model to adaptively scale the contribution of each IC map. This mimics a form of neural gain control, potentially introducing structured feature reweighting consistent with known neural gain modulation by enabling the model to emphasize or suppress distributed sources, analogous to dynamic reweighting in cortical circuits. While performance gains were modest, BrainSERF consistently achieved higher subject identification accuracy compared to the baseline.

BrainSurfGCN focused on computational efficiency, achieving more than 15 times reduction in trainable parameters without sacrificing performance. BrainSurfCNN and BrainSERF apply spherical convolution using parameterized differential operators that rely on four mesh representations: the original mesh, gradients along the x and y surface directions, and the Laplacian. While effective, this approach increases memory and compute demands by requiring multiple large tensors to be stored and processed at each layer. In contrast, our new BrainSurfGCN model uses a graph-based message-passing framework that represents the cortical surface as a node-edge graph and updates vertex features via a shared learnable weight matrix Θ. This design eliminates the need to explicitly compute or store differential operators, while still capturing local one-hop information propagation like spherical convolutions. The result is a dramatic reduction in model size (from ~100 to <1 MB) and training time (from ~26 to ~6 h), without sacrificing predictive performance. This efficiency makes BrainSurfGCN well-suited for resource-constrained environments, rapid prototyping, and potential clinical deployment, where model size and training speed are critical considerations. Although all models achieved comparable prediction accuracy, BrainSurfGCN substantially reduced computational cost, highlighting how graph structural designs may offer a scalable and efficient approach to individualized brain modeling.

4.2 | Data Requirements and Model Scalability

A critical consideration in connectome fingerprinting approaches is the amount of data required to train predictive models. CNN-based models, including those evaluated in this study, typically involve a large number of trainable parameters and therefore benefit from large-scale datasets. In our case, over 900 subjects were used for training, which may limit the applicability of such models in smaller or more specialized datasets. In contrast, prior work using regression and penalized regression approaches has demonstrated strong predictive performance and subject-level specificity even with relatively modest sample sizes (Osher et al. 2019; Tavor et al. 2016; Tobyne et al. 2018). These methods are therefore particularly attractive in data-limited settings.

Recent work has further highlighted alternative modeling strategies, including mechanistic models such as activity flow mapping (Cole et al. 2016), as well as deep learning approaches that directly model spatial structure (Ngo et al. 2022). These approaches highlight that prediction performance depends not only on model complexity but also on how connectivity information is represented and utilized. Within this broader landscape, BrainSurfGCN can be viewed as a middle ground between high-capacity deep learning models and more interpretable, low-parameter approaches. Notably, the reduction in model parameters in BrainSurfGCN is substantial (Table 1), suggesting a meaningful shift in the trade-off between model capacity and scalability. By incorporating topology-aware structural constraints, it reduces model complexity while maintaining comparable predictive performance to CNN-based models. This suggests potential improvements in data efficiency and scalability, making it a promising alternative when large training datasets are not available. Taken together, these findings highlight that model choice should be guided not only by predictive performance but also by data availability, computational cost, and the degree of structural inductive bias incorporated into the model.

4.3 | Positioning CNN-Based Models Within Connectome Fingerprinting Approaches

A growing body of work has demonstrated that task-evoked brain activity can be predicted from patterns of structural or functional connectivity using a range of modeling approaches (Cole et al. 2016; Ngo et al. 2022; Tavor et al. 2016). These approaches differ substantially in their underlying assumptions and modeling strategies. Regression-based methods provide interpretable mappings between connectivity features and task activation, and have been shown to capture strong subject-level specificity. Modeling approaches such as activity flow mapping (Cole et al. 2016) provide an explicit account of how task-evoked activity may arise from the propagation of activity over intrinsic connectivity networks. In contrast, CNN-based models (Ngo et al. 2022) are primarily optimized to reconstruct spatially structured activation patterns, often capturing spatial distributions effectively, but without explicitly modeling the underlying network topology. Prior work by Ngo et al. (2022) showed that CNN-based models achieve performance broadly comparable to both group-average baselines and linear regression approaches in terms of Dice AUC, with only modest differences across methods. However, these differences are relatively small, indicating that a substantial portion of task-evoked activation patterns can already be captured by simpler models.

Importantly, predictive performance across all approaches is fundamentally constrained by the reliability of the underlying data, as reflected by repeat measurements, which serve as a practical noise ceiling. Notably, the gap between model predictions and repeat data is substantially larger than the gap between different modeling approaches, underscoring the dominant role of intrinsic variability in limiting prediction accuracy. Within this broader landscape, our work focuses on how architectural design choices influence the trade-off between predictive performance, computational efficiency, and representational structure. Specifically, BrainSERF introduces channel-wise modulation mechanisms that enable adaptive feature reweighting, while BrainSurfGCN explicitly incorporates cortical topology to constrain information flow.

Unlike conventional CNN-based approaches that rely on local convolutional operations without explicitly modeling networks, BrainSurfGCN incorporates connectivity-informed message passing to constrain information flow. This design enables the model to capture interactions between distributed brain regions while maintaining comparable predictive performance with reduced computational cost. These findings suggest that embedding structured priors into model architecture offers an effective strategy for improving efficiency and interpretability in rest-to-task prediction.

4.4 | Impact of Input Representation and Dimensionality

Moreover, we examined the role of representational dimensionality by varying the number of ICA components used as model input. Contrary to prior findings that higher ICA dimensionality can improve prediction performance (Tavor et al. 2016; Tobyne et al. 2018), we observed that performance remained relatively stable across a wide range of component numbers. This suggests that predictive information may saturate at relatively low-dimensional representations in this setting.

One possible explanation is that the model architectures already capture relevant spatial structure through learned filters, reducing the marginal benefit of increasing input dimensionality. Alternatively, increasing ICA dimensionality may introduce redundant or noisy features that do not contribute additional predictive signal. It is also possible that prediction performance is fundamentally constrained by the reliability of task contrasts, such that improvements in input representation yield diminishing returns beyond a certain point. This may also indicate that rest-to-task prediction is limited more by the information content of the input data than by its dimensionality.

4.5 | Understanding Variability in Model Performance

Our second line of investigation focused on understanding why model performance varies across task contrasts and individuals. We observed that prediction accuracy was influenced by task engagement, data quality, and inter-subject variability.

First, we found preliminary evidence that poorer model predictions were associated with lower task accuracy and slower reaction times, particularly in tasks requiring sustained attention or complex reasoning. 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.

Notably, the observed relationships between prediction accuracy and behavioral performance were modest in magnitude. This is consistent with prior work showing that task-fMRI measures have limited test–retest reliability, and that behavioral performance reflects multiple cognitive processes that are only partially captured by large-scale activation patterns. Additionally, variability in attention and engagement during both resting-state and task scans may further attenuate these relationships. These findings should therefore be interpreted as preliminary indicators of sources of variability, rather than strong predictive links. This further suggests that variability in prediction performance may be driven more by underlying neural and behavioral variability than by model differences alone.

Second, we investigated why delta contrast maps (i.e., condition A vs. condition B) yielded lower prediction and test–retest reliability compared to contrasts versus baseline. An exploratory analysis of repeat error and tSNR maps revealed that delta contrasts showed greater overlap between signal dropout regions (low tSNR) and areas of high prediction error, particularly in contrasts involving subtle task manipulations (e.g., Math vs. Story, Punish vs. Reward). This suggests that delta contrasts may reflect noisier or less stable signal differences, making them more challenging targets for prediction models. Future studies should consider modeling task contrasts with varying levels of reliability separately or weighting voxels based on tSNR-informed confidence.

Third, we examined the role of inter-subject variability in shaping prediction performance. As illustrated in Figure 9, the degree of spatial consistency varied substantially across task contrasts. Regions associated with more stimulus-driven processing, such as visual areas, showed relatively high consistency across subjects, with activation patterns concentrated in similar locations. In contrast, less predictable contrasts, particularly delta contrasts, exhibited more diffuse and spatially variable activation patterns across individuals. This variability was also reflected in repeat measurements, indicating that it is not specific to model predictions. Together, these findings suggest that inter-subject variability in cortical activation, and the extent to which specific regions are consistently engaged across individuals, plays a key role in determining prediction performance.

4.6 | Limitations and Future Direction

Our study has several limitations, spanning both methodological design and generalization, that open avenues for future research: (1) Architectural scope: While BrainSERF and BrainSurfGCN introduced efficient design changes, the performance gains were incremental. New architectural innovations, such as equivariant GNNs, spectral attention, or hybrid encoding-decoding frameworks (Kondor 2025), may yield greater improvements in both accuracy and interpretability. (2) Generalization across mesh structures: BrainSurfGCN was evaluated on a fixed cortical mesh; however, its flexibility to operate on graphs of varying resolution remains untested. Leveraging coarser meshes, as done in the original BrainSurfCNN, may offer further speedups with minimal performance loss. (3) Zero-shot task generalization: Our models were trained to predict 47 fixed task contrasts. Extending these models to unseen tasks remains an open challenge and may require task embedding approaches or transfer learning on task instructions. (4) Group IC derivation: Resting-state features were constructed using group-level IC maps provided by HCP, which may include some overlap with test subjects. Although this risk is minimal given the dataset size, future work

should evaluate whether using subject-specific or held-out group ICs changes prediction fidelity. (5) Reliability of delta contrasts: Lower prediction performance for delta contrasts likely reflects their reduced reliability and higher noise sensitivity, a limitation consistently observed across prior work (Tavor et al. 2016; Tripathi and Somers 2023). This suggests that performance limits are driven more by intrinsic variability than by model capacity. (6) Cross-dataset generalization: All models were evaluated on HCP, limiting assessment of generalizability across datasets with different acquisition and population characteristics (Ngo et al. 2022; Tik et al. 2023). (7) Family structure: The HCP dataset includes related individuals, and although standard splits were used, potential dependencies between training and test sets may slightly inflate performance estimates.

4.7 | Toward Efficient and Scalable Predictive Modeling

Together, our findings highlight the importance of structurally informed modeling choices that enable efficient and scalable prediction of task-evoked brain activity from resting-state connectivity. In particular, BrainSurfGCN demonstrates that comparable predictive performance can be achieved with substantially reduced model complexity, emphasizing the value of parameter-efficient architectures for large-scale neuroimaging applications.

These results further support the idea that resting-state connectivity contains latent signatures of task-evoked activity, and that appropriately designed models can extract these signals in a subject-specific manner.

Moving forward, linking these predictive representations to behavior, trait variability, and clinical outcomes will be essential for translating these approaches into practical neuroscience and clinical settings. In addition, improving data quality, contrast reliability, and representation design will be critical for enhancing the robustness and generalizability of future models.

Author Contributions

Soren J. Madsen: conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing of original draft, reviewing and editing, and visualization. **Young-Eun Lee:** methodology, software, validation, reviewing and editing, and visualization. **Shaun K. L. Quah:** reviewing and editing. **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.

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.

Funding

This work was supported by the National Institute of Mental Health (MH127608) and Stanford Maternal and Child Health Research Institute, Faculty Scholar Award.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All code used for this article is made publicly available on GitHub at https://github.com/braindynamicslab/dl-task-contrast-prediction. The data that support the findings of this study are available in Human Connectome Project at https://www.humanconnectome.org/study/hcp-young-adult/document/hcp-young-adult-2025-release. These data were derived from the following resources available in the public domain: https://www.humanconnectome.org/study/hcp-young-adult/docume, https://www.humanconnectome.org/study/hcp-young-adult/document/hcp-young-adult-2025-release.

Endnotes

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

References

Barch, D. M., G. C. Burgess, M. P. Harms, et al. 2013. "Function in the Human Connectome: Task-Fmri and Individual Differences in Behavior." *NeuroImage* 80: 169–189.

Bernstein-Eliav, M., and I. Tavor. 2024. "The Prediction of Brain Activity From Connectivity: Advances and Applications." *Neuroscientist* 30, no. 3: 367–377.

Chow, W. W., A.-K. Seghouane, and M. L. Seghier. 2025. "A Statistical Characterization of Dynamic Brain Functional Connectivity." *Human Brain Mapping* 46, no. 2: e70145.

Cole, M. W., T. Ito, D. S. Bassett, and D. H. Schultz. 2016. "Activity Flow Over Resting-State Networks Shapes Cognitive Task Activations." *Nature Neuroscience* 19, no. 12: 1718–1726.

Cuthbert, B. N., and T. R. Insel. 2013. "Toward the Future of Psychiatric Diagnosis: The Seven Pillars of Rdoc." *BMC Medicine* 11, no. 1: 126.

Fey, M., and J. E. Lenssen. 2019. Fast Graph Representation Learning With PyTorch Geometric. ICLR Workshop on Representation Learning on Graphs and Manifolds.

Glasser, M. F., S. N. Sotiropoulos, J. A. Wilson, et al. 2013. "The Minimal Preprocessing Pipelines for the Human Connectome Project." *NeuroImage* 80: 105–124.

Gonzalez-Castillo, J., C. W. Hoy, D. A. Handwerker, et al. 2015. "Tracking Ongoing Cognition in Individuals Using Brief, Whole-Brain Functional Connectivity Patterns." *Proceedings of the National Academy of Sciences* 112, no. 28: 8762–8767. https://doi.org/10.1073/pnas.1501242112.

Hearne, L. J., R. D. Mill, B. P. Keane, G. Repovš, A. Anticevic, and M. W. Cole. 2021. "Activity Flow Underlying Abnormalities in Brain Activations and Cognition in Schizophrenia." *Science Advances* 7, no. 29: eabf2513. https://doi.org/10.1126/sciadv.abf2513.

Hu, J., L. Shen, and G. Sun. 2018. "Squeeze-and-Excitation Networks." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7132–7141. IEEE.

Insel, T. R. 2014. "The Nimh Research Domain Criteria (Rdoc) Project: Precision Medicine for Psychiatry." *American Journal of Psychiatry* 171, no. 4: 395–397.

Ioffe, S., and C. Szegedy. 2015. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift." *International Conference on Machine Learning* 37: 448–456.

Jiang, C., J. Huang, K. Kashinath, et al. 2019. Spherical Cnns on Unstructured Grids. arXiv Preprint arXiv:1901.02039.

Jiang, R., N. Zuo, J. M. Ford, et al. 2020. "Task-Induced Brain Connectivity Promotes the Detection of Individual Differences in Brain-Behavior Relationships." *NeuroImage* 207: 116370. https://doi.org/10.1016/j.neuroimage.2019.116370.

Jones, O. P., N. Voets, J. Adcock, R. Stacey, and S. Jbabdi. 2017. "Resting Connectivity Predicts Task Activation in Pre-Surgical Populations." *NeuroImage: Clinical* 13: 378–385.

Kingma, D. P., and J. Ba. 2014. Adam: A Method for Stochastic Optimization. arXiv Preprint arXiv:1412.6980.

Kipf, T. N., and M. Welling. 2016. Semi-Supervised Classification With Graph Convolutional Networks. arXiv Preprint arXiv:1609.02907.

Kondor, R. 2025. "The Principles Behind Equivariant Neural Networks for Physics and Chemistry." *Proceedings of the National Academy of Sciences* 122, no. 41: e2415656122.

Liu, R., J. Lehman, P. Molino, et al. 2018. "An Intriguing Failing of Convolutional Neural Networks and the Coordconv Solution." *Advances in Neural Information Processing Systems* 31: 9628–9639.

Milletari, F., N. Navab, and S.-A. Ahmadi. 2016. "V-net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation." In *Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV)*, 565–571. IEEE.

Ngo, G. H., M. Khosla, K. Jamison, A. Kuceyeski, and M. R. Sabuncu. 2022. "Predicting Individual Task Contrasts From Resting-State Functional Connectivity Using a Surface-Based Convolutional Network." *NeuroImage* 248: 118849. https://doi.org/10.1016/j.neuroimage.2021.118849.

Osher, D. E., J. A. Brissenden, and D. C. Somers. 2019. "Predicting an Individual's Dorsal Attention Network Activity From Functional Connectivity Fingerprints." *Journal of Neurophysiology* 122, no. 1: 232–240.

Osher, D. E., R. R. Saxe, K. Koldewyn, J. D. Gabrieli, N. Kanwisher, and Z. M. Saygin. 2016. "Structural Connectivity Fingerprints Predict Cortical Selectivity for Multiple Visual Categories Across Cortex." *Cerebral Cortex* 26, no. 4: 1668–1683.

Paszke, A., S. Gross, F. Massa, et al. 2019. "Pytorch: An Imperative Style, High-Performance Deep Learning Library." *Advances in Neural Information Processing Systems* 32: 8026–8037.

Quah, S., B. Jo, C. Geniesse, et al. 2025. "A Data-Driven Latent Variable Approach to Validating the Research Domain Criteria Framework." *Nature Communications* 16, no. 1: 830.

Quah, S., S. Madsen, S. Pirzada, et al. 2025. Refining Rdoc Using Individual-Level Task Fmri Factor Models Reveals Reproducible Brain-Wide Motifs. bioRxiv, 2025-10.

Ronneberger, O., P. Fischer, and T. Brox. 2015. "U-net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015." In *Proceedings of the 18th International Conference, October 5-9, 2015, Proceedings, Part III 18*, 234–241. Springer.

Savage, H. S., P. C. Mulders, P. F. Van Eijndhoven, et al. 2024. "Dissecting Task-Based Fmri Activity Using Normative Modelling: An Application to the Emotional Face Matching Task." *Communications Biology* 7, no. 1: 888.

Seghier, M. L., and C. J. Price. 2016. "Visualising Inter-Subject Variability in Fmri Using Threshold-Weighted Overlap Maps." *Scientific Reports* 6, no. 1: 20170.

Serin, E., K. Ritter, G. Schumann, T. Banaschewski, A. Marquand, and H. Walter. 2025. "Generating Synthetic Task-Based Brain Fingerprints for Population Neuroscience Using Deep Learning." *Communications Biology* 8, no. 1: 1572.

Smith, S. M., C. F. Beckmann, J. Andersson, et al. 2013. "Resting-State Fmri in the Human Connectome Project." *NeuroImage* 80: 144–168.

Smith, S. M., P. T. Fox, K. L. Miller, et al. 2009. "Correspondence of the Brain's Functional Architecture During Activation and Rest." *Proceedings of the National Academy of Sciences* 106, no. 31: 13040–13045.

Smith-Collins, A. P., K. Luyt, A. Heep, and R. A. Kauppinen. 2015. "High Frequency Functional Brain Networks in Neonates Revealed by Rapid Acquisition Resting State Fmri." *Human Brain Mapping* 36, no. 7: 2483–2494.

Sui, J., T. Adali, G. D. Pearlson, and V. D. Calhoun. 2009. "An Ica-Based Method for the Identification of Optimal FMRI Features and Components Using Combined Group-Discriminative Techniques." *NeuroImage* 46, no. 1: 73–86.

Tavor, I., O. P. Jones, R. B. Mars, S. Smith, T. Behrens, and S. Jbabdi. 2016. "Task-Free MRI Predicts Individual Differences in Brain Activity During Task Performance." *Science* 352, no. 6282: 216–220.

Tik, N., S. Gal, A. Madar, T. Ben-David, M. Bernstein-Eliav, and I. Tavor. 2023. "Generalizing Prediction of Task-Evoked Brain Activity Across Datasets and Populations." *NeuroImage* 276: 120213.

Tik, N., A. Livny, S. Gal, et al. 2021. "Predicting Individual Variability in Task-Evoked Brain Activity in Schizophrenia." *Human Brain Mapping* 42, no. 12: 3983–3992.

Tobyne, S. M., D. C. Somers, J. A. Brissenden, S. W. Michalka, A. L. Noyce, and D. E. Osher. 2018. "Prediction of Individualized Task Activation in Sensory Modality-Selective Frontal Cortex With 'Connectome Fingerprinting'." *NeuroImage* 183: 173–185.

Tripathi, V., and D. C. Somers. 2023. "Predicting an Individual's Cerebellar Activity From Functional Connectivity Fingerprints." *NeuroImage* 281: 120360.

Van Essen, D. C., S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, and K. Ugurbil. 2013. "The Wu-Minn Human Connectome Project: An Overview (Mapping the Connectome)." *NeuroImage* 80: 62–79. https://doi.org/10.1016/j.neuroimage.2013.05.041.

Wang, J., P. Lv, H. Wang, and C. Shi. 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.

Zhang, Y., L. Tetrel, B. Thirion, and P. Bellec. 2021. "Functional Annotation of Human Cognitive States Using Deep Graph Convolution." *NeuroImage* 231: 117847.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Figure S1:** 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. **Figure S2:** Thresholded task activation maps for 7 task contrasts using threshold of 25% and IC of 15. We compare ground truth, repeat, group average, and models' predicted activation for an individual (subject 917255). The thresholded activation maps on the right hemisphere (lateral view) represent, respectively, Language: Story, Relational Processing: Relational, Social Cognition: Theory of Mind, Emotion: Faces, Working Memory: 2-Back, Motor: Average, and Gambling: Reward. Ground truth (target), repeat scans, group average, and predictions from BrainSurfCNN, BrainSERF, and Brain-SurfGCN are presented with colors indicating

ground truth, model prediction, and overlap. Dice scores are displayed beneath each row, quantifying prediction fidelity across thresholds. The highest Dice value within each column is bolded to highlight the best-performing models. **Figure S3:** Thresholded task activation maps for 7 task contrasts using threshold of 25% and IC of 25. We compare ground truth, repeat, group average, and models' predicted activation for an individual (subject 917255). The thresholded activation maps on the right hemisphere (lateral view) represent, respectively, Language: Story, Relational Processing: Relational, Social Cognition: Theory of Mind, Emotion: Faces, Working Memory: 2-Back, Motor: Average, and Gambling: Reward. Ground truth (target), repeat scans, group average, and predictions from BrainSurfCNN, BrainSERF, and Brain-SurfGCN are presented with colors indicating ground truth, model prediction, and overlap. Dice scores are displayed beneath each row, quantifying prediction fidelity across thresholds. The highest Dice value within each column is bolded to highlight the best-performing models. **Figure S4:** Thresholded task activation maps for 7 task contrasts using threshold of 25% and IC of 100. We compare ground truth, repeat, group average, and models' predicted activation for an individual (subject 917255). The thresholded activation maps on the right hemisphere (lateral view) represent, respectively, Language: Story, Relational Processing: Relational, Social Cognition: Theory of Mind, Emotion: Faces, Working Memory: 2-Back, Motor: Average, and Gambling: Reward. Ground truth (target), repeat scans, group average, and predictions from BrainSurfCNN, BrainSERF, and Brain-SurfGCN are presented with colors indicating ground truth, model prediction, and overlap. Dice scores are displayed beneath each row, quantifying prediction fidelity across thresholds. The highest Dice value within each column is bolded to highlight the best-performing models. **Figure S5:** Subject identification accuracy across all task contrasts. **Figure S6:** Dice scores of all models across all task contrasts. **Table S1:** Dice AUC Across All Language, Relational, Social, Emotional, and Working Memory Tasks. **Table S2:** Dice AUC Across All Motor and Gambling Tasks. **Table S3:** 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. **Table S4:** Spatial correlation of percentage of subjects showing top 10% activation between predicted and ground truth contrast sets associated with motor and gambling tasks. **Figure S7:** Examples of unthresholded task activation for Language: Math-Story. We compare ground truth (target), repeat, model-predicted activation, and group average for an individual (subject 917255). **Figure S8:** Examples of unthresholded task activation for Emotion: Face. We compare ground truth (target), repeat, model-predicted, and group average activation for an individual (subject 917255). **Figure S9:** Examples of unthresholded task activation for Working Memory: 0-Back. We compare ground truth (target), repeat, model-predicted activation, and group average for an individual (subject 917255). **Figure S10:** Examples of unthresholded task activation for Motor: Left Hand-Average. We compare ground truth (target), repeat, model-predicted activation, and group average for an individual (subject 917255). **Figure S11:** Examples of unthresholded task activation for Language: Math-Story. We compare ground truth (target), repeat, model-predicted activation, and group average for an individual (subject 103818). **Figure S12:** Examples of unthresholded task activation for Social Cognition: Theory of Mind. We compare ground truth (target), repeat, model-predicted activation, and group average for an individual (subject 103818). **Figure S13:** Examples of unthresholded task activation for Emotion: Face. We compare ground truth (target), repeat, model-predicted activation, and group average for an individual (subject 103818). **Figure S14:** Examples of unthresholded task activation for Working Memory: 0-Back. We compare ground truth (target), repeat, model-predicted activation, and group average for an individual (subject 103818). **Figure S15:** Examples of unthresholded task activation for Motor: Left Hand-Average. We compare ground truth (target), repeat, model-predicted activation, and group average for an individual (subject 103818). **Figure S16:** Correlation difference for above- and below-average task accuracy groups across all task contrasts (BrainSurfGCN). Violin plots show the distribution of correlation difference for subjects with above-average (red) and below-average (blue) behavioral performance across all task contrasts with an available accuracy metric. After Benjamini-Hochberg FDR correction, only Working Memory: 2-Back Body and Working Memory: 2-Back Place exhibited significant group differences. **Figure S17:** Spatial overlap between rest-to-task prediction error and resting-state tSNR across all task contrasts. Dice AUC between top-t vertices of the absolute test-retest error map and the negative tSNR map (threshold sweep $t \in [0.05, 0.50]$), shown for all 47 HCP contrasts sorted in descending order. Delta contrasts (red) and single-condition contrasts (green); top-5 outlined in bold. Delta contrasts show significantly higher overlap (mean 0.108 vs 0.088; permutation $p = 0.001$, Cliff's $\delta = 0.48$).