

BrainMT: A Hybrid Mamba-Transformer Architecture for Modeling Long-Range Dependencies in Functional MRI Data

Arunkumar Kannan¹, Martin A. Lindquist², and Brian Caffo²

¹ Dept. of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, USA

² Dept. of Biostatistics, Johns Hopkins University, Baltimore, USA
akannan7@jhu.edu

Abstract. Recent advances in deep learning have made it possible to predict phenotypic measures directly from functional magnetic resonance imaging (fMRI) brain volumes, sparking significant interest in the neuroimaging community. However, existing approaches, primarily based on convolutional neural networks or transformer architectures, often struggle to model the complex relationships inherent in fMRI data, limited by their inability to capture long-range spatial and temporal dependencies. To overcome these shortcomings, we introduce BrainMT, a novel hybrid framework designed to efficiently learn and integrate long-range spatiotemporal attributes in fMRI data. Our framework operates in two stages: (1) a bidirectional Mamba block with a temporal-first scanning mechanism to capture global temporal interactions in a computationally efficient manner; and (2) a transformer block leveraging self-attention to model global spatial relationships across the deep features processed by the Mamba block. Extensive experiments on two large-scale public datasets, UKBioBank and the Human Connectome Project, demonstrate that BrainMT achieves state-of-the-art performance on both classification (sex prediction) and regression (cognitive intelligence prediction) tasks, outperforming existing methods by a significant margin. Our code and implementation details are available at [link](#).

Keywords: fMRI · Mamba · Transformers · Phenotypic Prediction

1 Introduction

The human brain is a complex, dynamic system whose intrinsic spatiotemporal organization can be investigated through functional connectivity, often modeled using functional magnetic resonance imaging (fMRI) [21] data. In recent years, there has been significant interest in building predictive models that explore the relationship between functional connectivity and cognition to enhance understanding of neurological and neuropsychiatric disorders [12,6,29,15].

Existing predictive approaches for fMRI data can be broadly categorized into correlation-based and voxel-based methods. Correlation-based methods typically

employ a two-step process to address the high dimensionality of 4D spatiotemporal fMRI data (three spatial dimensions and time). First, researchers use anatomical or functional parcellations or projections to reduce data dimensionality, followed by the computation of a subject-specific functional connectivity measure (typically a matrix). Data-driven models operate on these lower-dimensional matrices to predict a target variable associated with a downstream task. In the past decade, deep learning tools have been used as feature selection techniques in this pipeline, leveraging architectures such as convolutional neural networks [13,14], transformers [11,25], and graph neural networks [19]. However, these methods face two main drawbacks. 1) *Loss of spatial structure*: reducing 4D data to 2D matrices can discard important spatial information, especially when very-low dimensional parcellations are used [18]. 2) *Lack of consensus in parcellation strategies*: different parcellation schemes and connectivity metrics can yield varying model performance [1], motivating the need for predictive models that learn directly from raw volumetric fMRI data.

To address these limitations, recent studies have proposed using a voxel-based framework. Here preprocessed, voxel-level fMRI data are used as input in end-to-end deep learning pipelines. For instance, [23] proposed a transformer-centric architecture for predicting cognitive intelligence and classifying schizophrenia, while [16] developed a swin transformer-based framework for subject-specific phenotype prediction. Despite these advances, models are constrained by the quadratic complexity of transformers, forcing them to process only relatively short sequences of fMRI volumes (10 to 20 frames) and then aggregate predictions across multiple time-window batches. Given the relative sluggishness of the hemodynamics underlying fMRI signals [20], restricting the model to a few time frames may overlook predictive temporal dynamics. In fact, [16] shows that incorporating more time frames steadily improves model performance.

In this work, we propose BrainMT, a novel hybrid deep learning framework that addresses the above challenges by combining a bi-directional Mamba block with a global transformer module to capture the full spatiotemporal complexity of fMRI data. Drawing inspiration from recent Mamba architectures [17,26,10], we adopt a temporal-first scanning mechanism to efficiently handle extended fMRI sequences, mitigating the computational bottlenecks found in earlier approaches. This design enables BrainMT to preserve and model long-range temporal signals while leveraging a lightweight global transformer to learn spatial dependencies between brain regions. Through extensive experiments and ablation studies on two large-scale public datasets - UKBioBank (UKB) and the Human Connectome Project (HCP), we demonstrate that BrainMT outperforms existing methods and generalizes robustly across diverse tasks for improved phenotypic prediction in neuroimaging.

2 Methods

Figure. 1 provides an overview of our proposed BrainMT framework, which employs a cascaded design comprising three primary components: (1) a convolution

block; (2) a spatiotemporal Mamba block; and (3) a transformer block. Starting from fMRI data $\mathbf{X} \in \mathbb{R}^{T \times H \times W \times D}$ - where T represents the number of sampled volumes (i.e., one per TR; the time resolution of the signal), and H , W , and D denote the three spatial dimensions - we first partition each volume into partially overlapping patches of size $\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4}$. These patches are then projected into a C -dimensional embedding space using two convolution layers. The resulting embedded feature maps serve as inputs to the following sub-modules.

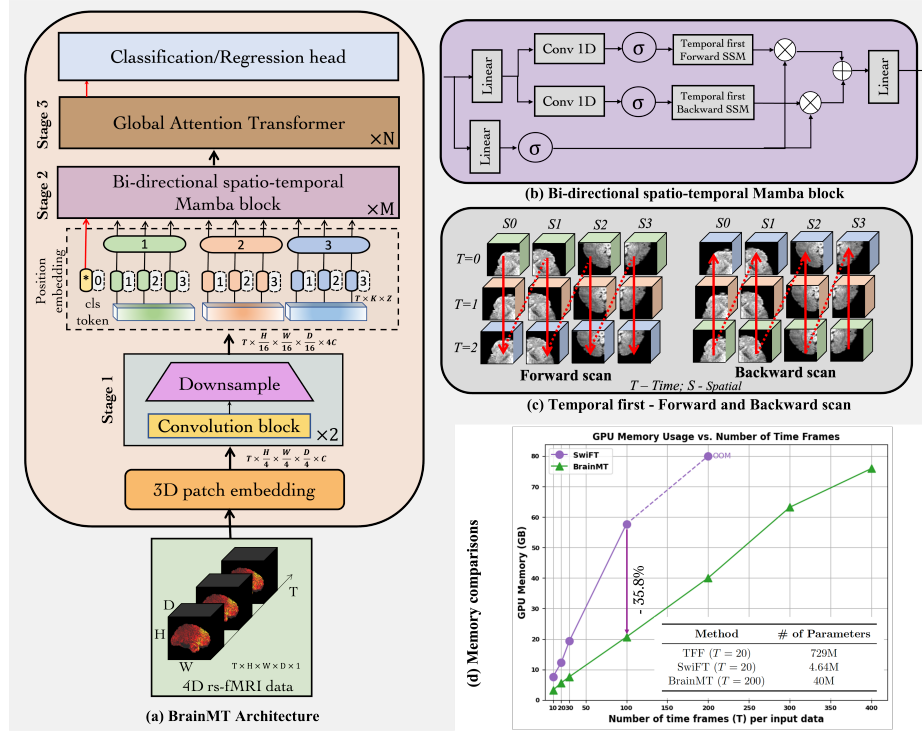


Fig. 1. Left: (a) Overall architecture of BrainMT. Right: (b) Structure of the bi-directional spatio-temporal Mamba block. (c) Scan mechanism. (d) Comparison of GPU memory usage and model parameters between BrainMT and SwiFT [16].

Convolution block: The convolution block acts as a hierarchical feature extractor for the embedded fMRI patches, generating multi-scale representations that capture coarse (high-resolution) and fine (low-resolution) details. In our framework, this block is implemented as a two-stage network, where each stage consists of a spatial convolution encoder followed by a downsampling operation. Formally, given embedded patches of size $\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4} \times C$, the convolution encoder outputs feature maps \mathbf{F}_i with dimensions $\frac{H}{4 \times 2^i} \times \frac{W}{4 \times 2^i} \times \frac{D}{4 \times 2^i}$ and channel

dimension $(C \times 2^i)$ for $i \in \{1, 2\}$, using the following residual configuration:

$$\begin{aligned}\hat{\mathbf{X}} &= \text{GELU}(\text{LN}(\text{Conv3D}(\mathbf{X}_{\text{in}}))) \\ \mathbf{X}_{\text{conv}} &= \text{LN}(\text{Conv3D}(\hat{\mathbf{X}})) + \mathbf{X}_{\text{in}}\end{aligned}$$

where GELU is the Gaussian Error Linear Unit activation function and LN denotes layer normalization.

Positional embedding: Next, we add learnable positional embeddings [3,4] to the convolution block output, as the following Mamba block relies on token positions to model long-range dependencies. Let $\mathbf{X}_{\text{conv}} \in \mathbb{R}^{T \times \frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times Z}$ with $Z = 4C$, be the convolution block’s output. First, each volumetric slice is flattened into a sequence of length $K = \frac{H}{16} \times \frac{W}{16} \times \frac{D}{16}$. We then add two learnable positional embeddings: a spatial embedding $\mathbf{P}_s \in \mathbb{R}^{1 \times K \times Z}$ and a temporal embedding $\mathbf{P}_t \in \mathbb{R}^{T \times 1 \times Z}$ to the sequence. Additionally, we prepend a learnable classification token $\mathbf{X}_{\text{cls}} \in \mathbb{R}^{1 \times 1 \times Z}$ to serve as a feature aggregator. This yields:

$$\mathbf{X}_{\text{out}} = [\mathbf{X}_{\text{cls}}, \mathbf{X}_{\text{conv}}] + \mathbf{P}_s + \mathbf{P}_t,$$

where $[\cdot, \cdot]$ denotes concatenation along the token dimension. Finally, we reshape \mathbf{X}_{out} from $\mathbb{R}^{T \times K \times Z}$ to $\mathbb{R}^{1 \times L \times Z}$ with $L = T \times K$, providing a longer context sequence for subsequent spatiotemporal modeling.

Spatio-temporal Mamba block: The Mamba block employs bi-directional selective state-space models (SSMs) to capture long-context representations from the input sequence. In what follows, we first outline the fundamentals of SSMs and their discretization in Mamba, and then describe the architecture that adapts Mamba to 4D spatiotemporal data.

(i) *SSM and Mamba preliminaries:* State-space models (SSMs), which originate from Kalman filters, are linear-time-invariant (LTI) systems that map a one-dimensional continuous input sequence $x(t) \in \mathbb{R}$ to a one-dimensional output sequence $y(t) \in \mathbb{R}$ via an internal hidden state $h(t) \in \mathbb{R}^N$, where N is the dimension of the hidden state. Formally,

$$h'(t) = \mathbf{A} h(t) + \mathbf{B} x(t), \quad (1)$$

$$y(t) = \mathbf{C} h(t), \quad (2)$$

where $\mathbf{A} \in \mathbb{R}^{N \times N}$ is the evolution matrix, and $\mathbf{B} \in \mathbb{R}^{N \times 1}$ and $\mathbf{C} \in \mathbb{R}^{1 \times N}$ are projection matrices. In order to handle discrete input data (e.g., images or fMRI sequences), Mamba [9] applies a Zero-Order Hold (ZOH) discretization to \mathbf{A} , \mathbf{B} , \mathbf{C} using a time-scale parameter Δ :

$$\bar{\mathbf{A}} = \exp(\Delta \mathbf{A}), \quad \bar{\mathbf{B}} = (\Delta \mathbf{A})^{-1} (\exp(\Delta \mathbf{A}) - \mathbf{I}) \Delta \mathbf{B}, \quad (3)$$

$$h(t) = \bar{\mathbf{A}} h(t-1) + \bar{\mathbf{B}} x(t), \quad y(t) = \mathbf{C} h(t). \quad (4)$$

Because SSMs are LTI, their parameters \mathbf{A} , \mathbf{B} , \mathbf{C} remain fixed for all time steps, which can limit contextual learning. To overcome this, Mamba introduces a

selective scan mechanism that make \mathbf{B} , \mathbf{C} , Δ input-dependent, thereby enabling the model to adaptively learn contextual relations from longer sequences.

(ii) *Architecture*: While the original Mamba formulation operates on 1D data (e.g., text), vision tasks require modeling spatial relationships. Hence, we adopt the Vision Mamba block [34], which incorporates bi-directional selective SSMs to capture spatiotemporal context in the data. Figure. 1b provides a schematic illustration. The input sequence is first linearly projected into two vectors, \mathbf{x} and \mathbf{z} , each of dimension d . To prevent tokens from being processed independently, a 1D convolution is applied to \mathbf{x} before feeding it into forward and backward selective SSM modules, producing $\mathbf{y}_{\text{forward}}$ and $\mathbf{y}_{\text{backward}}$. Both outputs are then gated by \mathbf{z} and summed to form the final output sequence. By processing the input in both directions and selectively adjusting the SSM parameters, the Vision Mamba block can learn rich spatiotemporal dependencies from fMRI data.

In addition, we adopt a *temporal-first scanning mechanism* as shown in Figure. 1c, where tokens are arranged so that time is treated as the leading dimension, followed by spatial dimensions. This design choice is motivated by neurobiological principles of coactivation in fMRI data, where consistent temporal correlations among brain regions are critical for identifying functional connectivity. By ensuring that the SSMs process the time dimension prior to spatial dimensions, the Vision Mamba block can more effectively leverage long-range temporal interactions alongside spatial dependencies from fMRI data.

Transformer block: The final stage of our framework leverages a multi-head self-attention module to capture global dependencies. Let $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{L \times Z}$ represent the query, key, and value matrices, where L is the sequence length and Z is the token dimensionality. The self-attention mechanism is:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q} \mathbf{K}^T}{\sqrt{d_{\text{head}}}}\right) \mathbf{V}, \quad (5)$$

where d_{head} is the dimension of each attention head. Although self-attention has a quadratic computational complexity in L , our initial convolution and down-sampling stages significantly reduce the sequence length, thereby making the transformer block tractable for longer sequences.

3 Experimental Setup and Results

Datasets and preprocessing: We applied our method to resting-state fMRI data from 6000 UKB participants [31] and 1075 HCP (S1200 release) participants [30], with UKB/HCP scan lengths of 490/1,200 volumes, spatial size of $91 \times 109 \times 91$ in MNI space, and class balances of female participants at 50.86% (UKB) and 51.16% (HCP). We used the preprocessed data provided with both datasets [8,2], which follows the ‘‘fMRI volume’’ pipeline (bias field reduction, skull stripping, cross-modality registration, and spatial normalization). To ensure stable network training, we applied global Z-score normalization, excluding background regions, which were filled with the minimum Z-score intensity

[23]. We split the data into 70% for training, 15% for validation, and 15% for testing. Our prediction targets were sex and cognitive intelligence scores: specifically the “cognitive function” composite scores from HCP and the “fluid intelligence/reasoning” scores from UKB. All regression targets were Z-score normalized. In correlation-based approaches, we parcellate the data using the HCP multimodal atlas [7] and compute Pearson correlations to capture functional connectivity between brain regions.

Implementation details: We implemented our model in PyTorch and trained it on NVIDIA L40S GPUs (48GB RAM). For all downstream tasks, we used the same BrainMT architecture with 2 convolution blocks, 12 Mamba blocks, and 8 transformer blocks in sequence. The final output is obtained by applying normalization to the \mathbf{X}_{cls} token, followed by an MLP head. We randomly selected 200 frames from resting-state fMRI as input (justified in Table 3) and used default Mamba hyperparameters for the spatiotemporal layer (state dimension: 16, expansion ratio: 2). Training used a distributed data-parallel strategy with AdamW under a cosine learning rate schedule over 20 epochs, with the first 5 as a linear warm-up. Default values were used for learning rate ($2\text{e-}4$), weight decay (0.05), and batch size (2), with early stopping based on validation loss. All hyperparameters were fixed using the validation set. For intelligence prediction, we minimized mean squared error (MSE) and evaluated using MSE, mean absolute error (MAE), and Pearson’s R. For sex classification, we optimized binary cross-entropy loss and evaluated using balanced accuracy (B.Acc), standard accuracy (Acc.), and the area under the receiver operating characteristic curve (AUROC).

Table 1. Comparisons with baseline models on HCP and UKB Datasets for cognitive intelligence prediction. Values are given with their standard deviations. Lower MSE and MAE, and higher R indicate better performance. Color convention: **best**, **2nd-best**.

Method	HCP							UKBioBank						
		MSE		MAE		R		MSE		MAE		R		
XG-Boost [5]		1.004	0.22	0.831	0.15	0.14	0.05	1.049	0.19	0.811	0.14	0.01	0.03	
BrainNetCNN [13]		0.981	0.21	0.799	0.10	0.21	0.02	1.003	0.23	0.801	0.16	0.01	0.01	
BrainGNN [19]		0.946	0.02	0.791	0.10	0.28	0.04	0.995	0.13	0.794	0.03	0.06	0.10	
BrainNetTF [11]		0.998	0.17	0.820	0.03	0.18	0.10	0.999	0.01	0.798	0.15	0.04	0.05	
TFF [23]		0.957	0.10	0.798	0.02	0.27	0.08	0.998	0.02	0.795	0.11	0.04	0.03	
SwiFT [16]		0.914	0.11	0.790	0.04	0.32	0.02	0.994	0.16	0.791	0.03	0.07	0.01	
BrainMT		0.835	0.02	0.741	0.01	0.41	0.03	0.932	0.03	0.773	0.01	0.24	0.02	

Quantitative results: To evaluate the effectiveness of BrainMT, we performed a comprehensive analysis against state-of-the-art methods representing both correlation-based and voxel-based approaches. For the correlation-based methods, we compared BrainMT with XGBoost [5] (our standard machine learning baseline), BrainNetCNN [13], BrainGNN [19], and BrainNetTF [11]. For the voxel-based methods, we used TFF [23] and SwiFT [16] as baselines. In each

Table 2. Comparisons with baseline models for sex classification. Higher Acc., B.Acc, and AUROC indicate better performance. Color convention: **best**, **2nd-best**

Method	HCP							UKBioBank						
	Acc.	B.Acc	AUROC					Acc.	B.Acc	AUROC				
XG-Boost [5]	68.43	2.37	67.86	3.35	73.2	2.43		79.15	1.38	78.27	1.37	86.3	0.42	
BrainNetCNN [13]	76.94	3.29	75.41	2.31	82.3	2.35		85.78	0.41	84.82	0.38	92.4	0.32	
BrainGNN [19]	84.73	1.22	84.26	0.98	90.1	0.73		89.85	1.28	89.71	0.21	95.8	1.14	
BrainNetTF [11]	82.87	2.19	81.55	3.16	89.3	2.09		87.92	1.23	87.38	1.15	95.1	0.71	
TFF [23]	92.94	1.17	92.40	1.19	97.1	2.08		96.11	0.13	95.46	0.28	99.3	0.12	
SwiFT [16]	93.06	1.08	92.61	1.26	97.5	1.95		97.45	0.11	97.69	0.05	99.4	0.13	
BrainMT	96.28	0.02	96.15	0.03	98.7	0.06		97.91	0.09	97.77	0.08	99.2	0.04	

case, we followed the original studies’ implementations, and all baselines were hyperparameter-tuned on the validation split. Quantitative results obtained from a repeated three-fold cross-validation are reported in Table. 1 for cognitive intelligence prediction and in Table. 2 for sex classification. On the intelligence prediction task, BrainMT consistently outperformed all baseline methods by a significant margin. Notably, most baseline methods yielded an MSE close to 1.0 in UKB (with targets normalized to have a variance of 1), suggesting that they mostly predict the sample mean. In contrast, BrainMT achieved a 6.23% reduction in MSE on the UKB dataset and a 8.75% reduction on the HCP dataset. For sex classification, BrainMT significantly outperformed all baselines on the HCP dataset and matched the performance of SwiFT on the UKB dataset. We further show that BrainMT is memory-efficient in Figure 1d, where GPU memory usage and model parameters are compared with SwiFT across varying frame counts, demonstrating that BrainMT is 35.8% more memory-efficient and maintains linear complexity in T .

Table 3. Ablation studies on different number of time frames (\mathcal{A}), architecture configurations (\mathcal{B}), number of Mamba (M) and Transformer (N) layers (\mathcal{C}), visual mamba blocks (\mathcal{D}) and predicting functional connectivity correlations (\mathcal{E}).

Exp	Configuration	HCP							UKBioBank						
		MSE	MAE	R					MSE	MAE	R				
\mathcal{A}	$T = 100$	0.861	0.19	0.765	0.14	0.38	0.11		0.966	0.12	0.783	0.08	0.20	0.05	
	$T = 300$	0.870	0.08	0.767	0.11	0.37	0.06		0.971	0.15	0.785	0.16	0.19	0.13	
\mathcal{B}	No Transformer	0.865	0.02	0.766	0.13	0.37	0.05		0.962	0.16	0.782	0.07	0.20	0.01	
	No Conv	0.849	0.17	0.754	0.09	0.39	0.11		0.949	0.13	0.778	0.18	0.21	0.03	
\mathcal{C}	No Conv & Transf.	0.857	0.06	0.759	0.15	0.38	0.04		0.955	0.15	0.779	0.12	0.20	0.08	
	Large (24M, 16N)	0.874	0.04	0.768	0.07	0.37	0.08		0.951	0.01	0.780	0.08	0.21	0.05	
\mathcal{D}	Small (6M, 4N)	0.883	0.14	0.772	0.10	0.36	0.16		0.965	0.04	0.783	0.07	0.20	0.09	
	VMamba [22]	0.891	0.38	0.790	0.20	0.24	0.13		0.968	0.29	0.783	0.28	0.15	0.19	
\mathcal{E}	MambaVision [10]	0.913	0.21	0.801	0.25	0.07	0.01		0.979	0.10	0.787	0.23	0.09	0.18	
	SwiFT	0.259	0.48	0.047	0.21	0.54	0.35		0.126	0.29	0.031	0.27	0.65	0.31	
\mathcal{E}	BrainMT	0.187	0.38	0.035	0.25	0.63	0.14		0.083	0.11	0.027	0.19	0.72	0.18	

Ablation studies: To assess the design choices of our framework, we conducted extensive ablation experiments $\mathcal{A} - \mathcal{E}$ on BrainMT for cognitive intelligence prediction, summarized in Table 3. Exp \mathcal{A} shows that our choice of $T = 200$ is optimal, likely because fewer frames yield an insufficient signal-to-noise ratio, whereas more frames increase the risk of overfitting. Exp \mathcal{B} validates our hybrid design, where convolution extracts local spatial features, Mamba captures sequential dynamics, and transformers encode global context. Exp \mathcal{C} finds 12 Mamba layers and 8 transformers to be the best trade-off between complexity and capacity. Exp \mathcal{D} shows that replacing the bi-directional vision Mamba block with alternatives (e.g., VMamba [22], MambaVision [10]) degrades performance, emphasizing the importance of causal convolution for preserving fMRI’s temporal order. Finally, Exp \mathcal{E} demonstrates that our model surpasses SwiFT in subject-level Pearson correlations by processing more frames at once, capturing sequential brain dynamics more effectively.

Clinical biomarkers: To identify the brain biomarkers underlying BrainMT’s phenotype predictions, we applied the Integrated Gradients (IG) algorithm [32] to assign importance scores to input features. As shown in Figure 2, the IG maps for cognitive intelligence reveal key contributions from regions within default mode network (DMN) and frontoparietal network (FPN), including posterior cingulate cortex (PCC), anterior cingulate cortex (ACC), precuneus (PCu), and cuneus (Cu). These areas are well-established in the literature for their roles in working memory, attention, decision-making, and visuospatial processing [28,24]. For sex prediction, the IG maps consistently highlight superior temporal gyrus (STG), middle frontal gyrus (MFG), and PCu - findings that align with previous neuroimaging studies on sex differences [27,33].

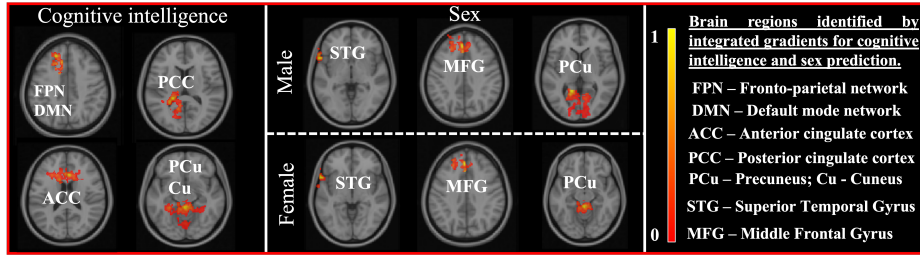


Fig. 2. Consistent interpretability maps from BrainMT model, highlighting regions identified by Integrated Gradients for predicting cognitive intelligence and sex in HCP.

4 Conclusion

In this work, we introduced BrainMT, a novel hybrid deep network that captures the complex long-range temporal and spatial dependencies in volumetric

fMRI data. By integrating a bi-directional Mamba block and a lightweight transformer block, BrainMT efficiently handles high-dimensional time series data and facilitates subject-level phenotypic characterizations. To the best of our knowledge, this is the first hybrid approach combining Mamba and transformer modules to manage extensive temporal coverage in volumetric rs-fMRI. This framework could serve as a general approach for prediction based on fMRI. In future work, we plan to incorporate self-supervised training on large-scale neuroimaging datasets to learn robust representations and further evaluate the model's generalizability to downstream tasks with limited data.

Acknowledgments. This work was supported by the National Institutes of Health grants R01 EB029977 (PI Caffo) and R01 EB026549 (PI Lindquist) from the National Institute of Biomedical Imaging and Bioengineering. This work used data from the Human Connectome Project, WU-Minn Consortium (PIs David Van Essen and Kamil Ugurbil; NIH grant 1U54 MH091657, funded by the 16 NIH Institutes and Centers supporting the NIH Blueprint for Neuroscience Research and the McDonnell Center for Systems Neuroscience at Washington University), as well as from UK Biobank (Project ID 33278), a major biomedical database.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abraham, A., Milham, M.P., Di Martino, A., Craddock, R.C., Samaras, D., Thirion, B., Varoquaux, G.: Deriving reproducible biomarkers from multi-site resting-state data: An autism-based example. *NeuroImage* **147**, 736–745 (2017)
2. Alfaro-Almagro, F., Jenkinson, M., Bangerter, N.K., Andersson, J.L., Griffanti, L., Douaud, G., Sotiropoulos, S.N., Jbabdi, S., Hernandez-Fernandez, M., Vallee, E., et al.: Image processing and quality control for the first 10,000 brain imaging datasets from uk biobank. *Neuroimage* **166**, 400–424 (2018)
3. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 6836–6846 (2021)
4. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: *ICML*. vol. 2, p. 4 (2021)
5. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. pp. 785–794 (2016)
6. Dsouza, N.S., Nebel, M.B., Crocetti, D., Robinson, J., Mostofsky, S., Venkataraman, A.: M-gcn: A multimodal graph convolutional network to integrate functional and structural connectomics data to predict multidimensional phenotypic characterizations. In: *Medical Imaging with Deep Learning*. pp. 119–130. PMLR (2021)
7. Glasser, M.F., Coalson, T.S., Robinson, E.C., Hacker, C.D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C.F., Jenkinson, M., et al.: A multi-modal parcellation of human cerebral cortex. *Nature* **536**(7615), 171–178 (2016)
8. 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.: The minimal

- preprocessing pipelines for the human connectome project. *Neuroimage* **80**, 105–124 (2013)
9. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
 10. Hatamizadeh, A., Kautz, J.: Mambavision: A hybrid mamba-transformer vision backbone. arXiv preprint arXiv:2407.08083 (2024)
 11. Kan, X., Dai, W., Cui, H., Zhang, Z., Guo, Y., Yang, C.: Brain network transformer. *Advances in Neural Information Processing Systems* **35**, 25586–25599 (2022)
 12. Kannan, A., Caffo, B., Venkataraman, A.: Gaming the brain: Investigating the cross-modal relationships between functional connectivity and structural features using generalized additive models. In: *International Workshop on Machine Learning in Clinical Neuroimaging*. pp. 166–175. Springer (2025)
 13. Kawahara, J., Brown, C.J., Miller, S.P., Booth, B.G., Chau, V., Grunau, R.E., Zwicker, J.G., Hamarneh, G.: Brainnetcnn: Convolutional neural networks for brain networks; towards predicting neurodevelopment. *NeuroImage* **146**, 1038–1049 (2017)
 14. Khosla, M., Jamison, K., Kuceyeski, A., Sabuncu, M.R.: Ensemble learning with 3d convolutional neural networks for functional connectome-based prediction. *NeuroImage* **199**, 651–662 (2019)
 15. Khosla, M., Jamison, K., Ngo, G.H., Kuceyeski, A., Sabuncu, M.R.: Machine learning in resting-state fmri analysis. *Magnetic resonance imaging* **64**, 101–121 (2019)
 16. Kim, P., Kwon, J., Joo, S., Bae, S., Lee, D., Jung, Y., Yoo, S., Cha, J., Moon, T.: Swift: Swin 4d fmri transformer. *Advances in Neural Information Processing Systems* **36**, 42015–42037 (2023)
 17. Li, K., Li, X., Wang, Y., He, Y., Wang, Y., Wang, L., Qiao, Y.: Videomamba: State space model for efficient video understanding. In: *European Conference on Computer Vision*. pp. 237–255. Springer (2024)
 18. Li, W., Lin, X., Chen, X.: Detecting alzheimer's disease based on 4d fmri: An exploration under deep learning framework. *Neurocomputing* **388**, 280–287 (2020)
 19. Li, X., Zhou, Y., Dvornek, N., Zhang, M., Gao, S., Zhuang, J., Scheinost, D., Staib, L.H., Ventola, P., Duncan, J.S.: Braingnn: Interpretable brain graph neural network for fmri analysis. *Medical Image Analysis* **74**, 102233 (2021)
 20. Lindquist, M.A.: The statistical analysis of fmri data. *Statistical science* (4), 439–464 (2008)
 21. Lindquist, M.A., Smith, B.B., Kannan, A., Zhao, A., Caffo, B.: Measuring the functioning human brain. *Annual Review of Statistics and Its Application* **12** (2024)
 22. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., Liu, Y.: VMamba: Visual state space model. In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems* (2024), <https://openreview.net/forum?id=ZgtLQQR1K7>
 23. Malkiel, I., Rosenman, G., Wolf, L., Hendler, T.: Self-supervised transformers for fmri representation. In: *International Conference on Medical Imaging with Deep Learning*. pp. 895–913. PMLR (2022)
 24. Menon, V.: Large-scale brain networks and psychopathology: a unifying triple network model. *Trends in cognitive sciences* **15**(10), 483–506 (2011)
 25. Nguyen, S., Ng, B., Kaplan, A.D., Ray, P.: Attend and decode: 4d fmri task state decoding using attention models. In: *Machine Learning for Health*. pp. 267–279. PMLR (2020)

26. Park, J., Kim, H.S., Ko, K., Kim, M., Kim, C.: Videomamba: Spatio-temporal selective state space model. In: European Conference on Computer Vision. pp. 1–18. Springer (2024)
27. Ryali, S., Zhang, Y., de Los Angeles, C., Supekar, K., Menon, V.: Deep learning models reveal replicable, generalizable, and behaviorally relevant sex differences in human functional brain organization. *Proceedings of the National Academy of Sciences* **121**(9), e2310012121 (2024)
28. Sestieri, C., Corbetta, M., Romani, G.L., Shulman, G.L.: Episodic memory retrieval, parietal cortex, and the default mode network: functional and topographic analyses. *Journal of Neuroscience* **31**(12), 4407–4420 (2011)
29. Smith, B.B., Zhao, Y., Lindquist, M.A., Caffo, B.: Regression models for partially localized fmri connectivity analyses. *Frontiers in Neuroimaging* **2**, 1178359 (2023)
30. 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.: Resting-state fmri in the human connectome project. *Neuroimage* **80**, 144–168 (2013)
31. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al.: Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine* **12**(3), e1001779 (2015)
32. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International conference on machine learning. pp. 3319–3328. PMLR (2017)
33. Weis, S., Patil, K.R., Hoffstaedter, F., Nostro, A., Yeo, B.T., Eickhoff, S.B.: Sex classification by resting state brain connectivity. *Cerebral cortex* **30**(2), 824–835 (2020)
34. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417* (2024)