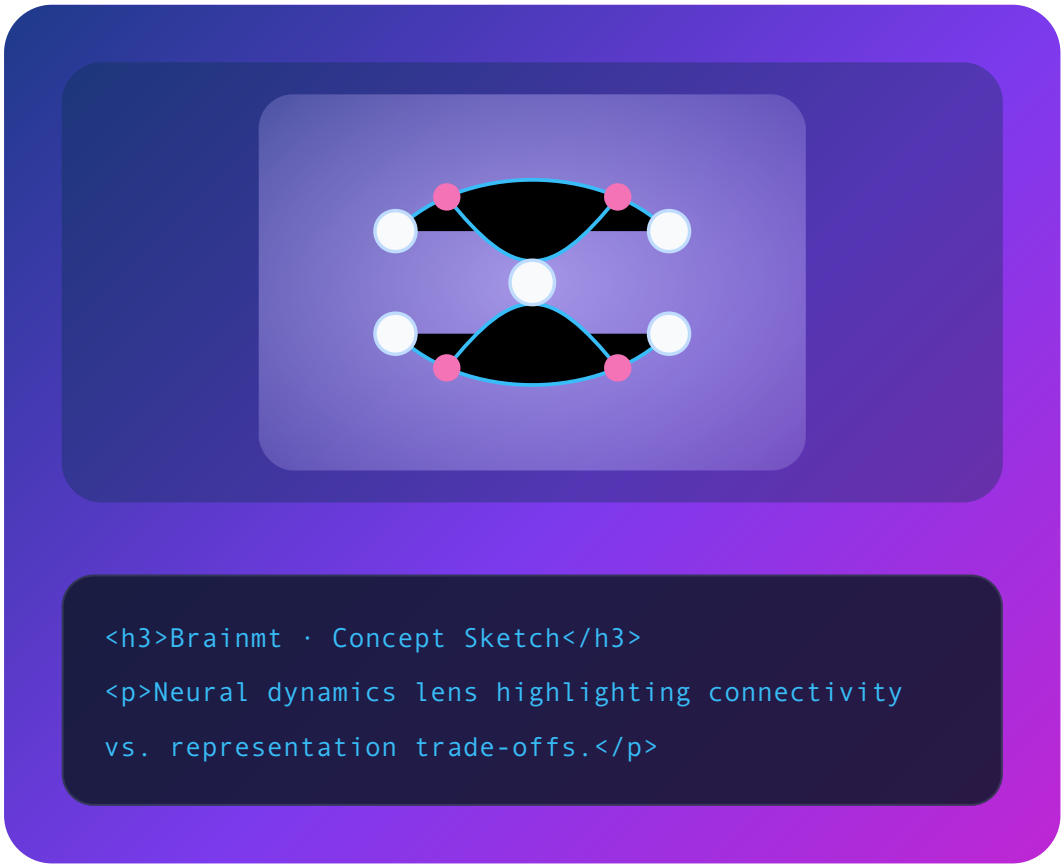

BrainMT: A Hybrid Mamba-Transformer Architecture for Modeling Long-Range Dependencies in Functional MRI Data

Authors: Arunkumar Kannan, Martin A. Lindquist, Brian Caffo
Year: Unknown (not clearly specified in extracted text)
Venue: Unknown (likely a neuroimaging / ML venue; not clearly specified in extracted text)

1. Classification

- **Domain Category:**
  - **Brain FM.** The paper develops a deep architecture specifically for resting-state functional MRI (rs-fMRI) and uses it to predict subject-level phenotypes (sex and cognitive intelligence) from whole-brain 4D volumes.
- **FM Usage Type:**
  - **Core FM development (brain-specific sequence model).** The work introduces a new hybrid architecture, BrainMT, that combines Mamba state-space models and transformers to handle long 4D fMRI sequences end-to-end. It is not a huge general-purpose foundation model in the GPT/CLIP sense, but it *does* propose a new, reusable backbone for a broad class of fMRI prediction tasks.
- **Key Modalities:**
  - Resting-state fMRI (4D volumetric time series: 3D brain volumes over time).
  - Phenotypic labels: sex (classification) and cognitive intelligence scores (regression).

---

2. One-Paragraph High-Level Summary (For a New Grad Student)

This paper proposes BrainMT, a deep learning architecture designed to model long-range spatial and temporal dependencies in resting-state fMRI data for subject-level phenotypic prediction. Instead of compressing fMRI into connectivity matrices or using short temporal windows, BrainMT operates directly on 4D voxel-wise volumes and can process much longer temporal sequences efficiently. The model uses a hierarchical convolutional block to extract local spatial features, a bidirectional spatiotemporal Mamba block (a modern state-space model) to capture long-range temporal and spatiotemporal patterns, and a lightweight transformer block to model global spatial relationships. The

authors evaluate BrainMT on large UK Biobank and Human Connectome Project datasets, predicting both sex and cognitive intelligence, and show it outperforms strong baselines including graph neural networks, transformer-based models, and the recent SwiFT voxel-wise transformer. Quantitative and ablation studies demonstrate that BrainMT achieves better accuracy with lower memory usage while benefiting from longer temporal context. Finally, interpretability analysis using Integrated Gradients highlights brain regions in default mode and frontoparietal networks that align with known neuroscience findings. Overall, the work is valuable to a new grad student because it showcases how modern sequence modeling ideas (Mamba + transformers) can be adapted to challenging 4D neuroimaging data and why long temporal context matters for fMRI-based prediction.

---

3. Problem Setup and Motivation

- **Scientific / practical problem:**
  - The goal is to predict subject-level phenotypes (sex and cognitive intelligence) from resting-state fMRI data.
  - Each subject has a 4D fMRI scan: a sequence of 3D brain volumes over time (volumes × height × width × depth).
  - The model should learn functional connectivity and spatiotemporal patterns directly from these volumetric time series, without relying on hand-crafted parcellations or connectivity matrices.
- **Why existing approaches are limited:**
  - **Correlation-based pipelines:**
    - They typically parcellate the brain into regions of interest (ROIs) and compute a region-by-region functional connectivity matrix (e.g., Pearson correlation).
    - This reduces dimensionality but can discard fine-grained spatial information, especially with coarse parcellations.

        - Performance can vary widely depending on the chosen parcellation and connectivity measure, leading to instability and lack of consensus.
    - **Voxel-based deep models:**
        - Recent transformer or CNN-based methods operate directly on voxel-level fMRI, but transformer-based models face quadratic complexity in sequence length.
        - As a result, they are often restricted to short sequences (e.g., 10–20 time frames), then aggregate predictions across multiple windows.
        - Because fMRI signals evolve relatively slowly (hemodynamics), limiting context to a few frames can miss important long-range temporal dynamics.
- **Why this is hard technically:**
    - fMRI is **high-dimensional**: 3D volumes with tens of thousands of voxels repeated hundreds to thousands of times per scan.
    - Spatiotemporal patterns are **long-range and structured**: activity in distant brain regions can be correlated across long time scales.
    - Direct transformer modeling over all voxels and time points is computationally heavy (quadratic in sequence length).
    - Models must balance **expressivity** (capturing complex patterns) with **efficiency** (memory and compute) to be practical on large datasets.

---

4. Data and Modalities

- **Datasets used:**
    - **UK Biobank (UKB):**
        - Resting-state fMRI data from about 6,000 participants.
        - Scan length: approximately 490 volumes per subject.
        - Spatial dimensions: 91 × 109 × 91 in MNI space.

        - Sex distribution: ~50.86% female.
    - **Human Connectome Project (HCP, S1200 release):**
        - Resting-state fMRI data from 1,075 participants.
        - Scan length: approximately 1,200 volumes per subject.
        - Spatial dimensions: 91 × 109 × 91 in MNI space.
        - Sex distribution: ~51.16% female.
- **Modalities:**
    - Primary modality: resting-state fMRI volumes (4D: time × 3D brain).
    - Targets:
        - **Cognitive intelligence scores:**
            - "Cognitive function" composite scores from HCP.
            - "Fluid intelligence/reasoning" scores from UKB.
        - **Sex:** binary labels for sex classification.
- **Preprocessing / representation:**
    - The authors use existing preprocessed fMRI data from both datasets, following "fMRI volume" pipelines that include:
        - Bias field reduction, skull stripping, cross-modality registration, and spatial normalization.
    - For modeling:
        - Global Z-score normalization is applied across brain voxels, excluding background regions; background voxels are filled with minimum Z-score intensity.
        - Each 3D volume is partitioned into partially overlapping 3D patches.
        - Patches are embedded via convolutional layers into a lower-dimensional feature space before being passed into the Mamba and transformer blocks.
    - For comparison with correlation-based approaches, the HCP multimodal atlas is used to parcellate the data and compute ROI–ROI Pearson correlation matrices.

5. Model / Foundation Model

BrainMT is a **hybrid architecture** combining convolutional neural networks, Mamba state-space models, and transformers to capture local, long-range temporal, and global spatial dependencies in fMRI.

- **Model type:**
  - A **hybrid deep sequence model** for spatiotemporal fMRI data, composed of:
    - 3D convolutional blocks (hierarchical feature extractor).
    - Bidirectional **Vision Mamba** blocks (selective state-space models) arranged with a temporal-first scanning mechanism.
    - A **multi-head self-attention transformer block** for global spatial modeling.
  - Overall, it can be viewed as a brain-specific backbone that blends linear-time sequence modeling (Mamba) with transformer attention.
- **Is it a new FM or an existing one?**
  - The architecture itself, **BrainMT**, is new.
  - It builds on existing ideas: Vision Mamba / Mamba-based state-space models and standard transformers.
  - The paper does not present large-scale general-purpose pretraining; instead, BrainMT is trained directly on the task objectives (sex and intelligence prediction) on large datasets.
  - So in FM terms, it is closer to a **new, reusable backbone** than a fully generic foundation model.
- **Key architectural components and innovations:**
  - **1. Convolution block (local spatial feature extractor):**
    - Takes fMRI volumes (X ^{T H W D}).
    - Each volume is split into partially overlapping 3D patches (downsampling the spatial resolution).

        - Two convolutional layers project patches into a C-dimensional embedding.
        - A two-stage convolutional encoder with downsampling forms multi-scale feature maps, capturing coarse and fine spatial details.
        - Residual convolutional units with GELU activations and layer normalization provide stable feature extraction.
    - **2. Positional embeddings and sequence construction:**
        - The output of the convolution block is a sequence of patch tokens per time step.
        - Learnable **spatial positional embeddings** and **temporal positional embeddings** are added to encode positions.
        - A learnable **classification token** (X_{}) is prepended as a global aggregator.
        - Tokens are reshaped into a long 1D sequence of length (L = T K) (where (K) is the number of patches), preparing them for spatiotemporal modeling.
    - **3. Spatiotemporal Mamba block (long-range context):**
        - Based on **selective state-space models (SSMs)**, originally inspired by Kalman-like systems.
        - Uses **Vision Mamba** with **bidirectional selective SSMs** to handle spatiotemporal context.
        - Mamba introduces input-dependent parameters (e.g., B, C, and time scales), enabling adaptive context modeling.
        - The **temporal-first scanning mechanism** arranges tokens so that time is the leading dimension, followed by spatial dimensions, which:
            - Emphasizes temporal continuity and long-range temporal correlations critical for fMRI.
            - Helps the SSM capture long sequences with linear-time complexity in sequence length.

        - Forward and backward selective SSMs process the sequence in both directions, and their outputs are gated and combined, enabling rich bidirectional context.
    - **4. Transformer block (global spatial relationships):**
        - After Mamba processing, a multi-head self-attention transformer operates on the sequence.
        - Attention is defined as usual ((Q,K,V) = (QK^/) V).
        - Because convolution and downsampling have already reduced the spatial resolution, the sequence length is shorter, making quadratic transformer attention tractable.
        - This block focuses on **global spatial interactions** among feature tokens, complementing the temporally oriented Mamba block.
    - **5. Output head:**
        - The final representation is taken from the normalized classification token (X_{}).
        - A multilayer perceptron (MLP) head is applied for downstream tasks:
            - Regression head for cognitive intelligence prediction.
            - Classification head for sex prediction.

- **Training setup (as described):**

Aspect	Description
Implementation	PyTorch; trained on NVIDIA L40S GPUs (48GB RAM)
Architecture depth	2 convolution blocks, 12 Mamba blocks, 8 transformer blocks
Input frames	200 frames per subject (chosen via ablation; trade-off between SNR and overfitting)

Aspect	Description
Mamba hyperparameters	State dimension 16, expansion ratio 2 (default Mamba settings)
Optimization	AdamW, cosine learning rate schedule over 20 epochs
Warm-up	First 5 epochs used for linear warm-up
Learning rate	2e-4 (default in experiments)
Weight decay	0.05
Batch size	2
Training strategy	Distributed data-parallel training
Early stopping	Based on validation loss
Regression objective	Mean squared error (MSE), evaluated with MSE, MAE, Pearson's R
Classification objective	Binary cross-entropy, evaluated with accuracy, balanced accuracy, AUROC

6. Multimodal / Integration Aspects (If Applicable)

- **Is this a multimodal integration paper?**
  - Not primarily. The core focus is on **single-modality** resting-state fMRI data.

    - The model does not fuse different data types (e.g., structural MRI, behavior, genetics) within the architecture; instead, it learns from volumetric fMRI alone and predicts phenotypes.

- **Integration aspects (within fMRI):**
    - BrainMT integrates **spatial and temporal information** from fMRI in a unified framework:
        - Convolution handles local spatial integration.
        - Mamba handles long-range temporal and spatiotemporal dependencies.
        - Transformers handle global spatial relationships across the entire brain.
    - This can be viewed as **intra-modality integration** (across time and space) rather than multimodal integration.
- **Relation to the integration baseline plan:**
    - The integration baseline plan emphasizes **late fusion across heterogeneous modalities**, CCA-based analysis, and disciplined evaluation.
    - BrainMT does not explicitly adopt late fusion or CCA-style strategies, as it operates on a single modality.
    - However, its design aligns with the principle of **preserving rich modality-specific signal** (here, fMRI) by avoiding aggressive parcellation and instead modeling voxel-level dynamics directly.
    - The evaluation uses solid metrics (MSE/MAE/R for regression; accuracy, balanced accuracy, AUROC for classification), which resonates with the plan's emphasis on robust, properly reported metrics.

Because the paper is not truly multimodal, the integration baseline plan is more of a conceptual backdrop here than a direct methodological influence.

7. Experiments and Results

- **Tasks / benchmarks:**
  - **Cognitive intelligence prediction (regression):**
    - Predicts continuous intelligence scores for subjects in HCP and UKB.
    - Evaluated with MSE, MAE, and Pearson's correlation R.
  - **Sex classification:**
    - Binary classification of sex for HCP and UKB participants.
    - Evaluated with accuracy, balanced accuracy, and AUROC.
  - **Additional analysis:**
    - Predicting functional connectivity correlations, comparing BrainMT with SwiFT on subject-level Pearson correlations.
    - Ablation studies on number of frames, architecture components, numbers of layers, and alternative Mamba variants.
    - Interpretability analyses to identify brain regions contributing to predictions.
- **Baselines:**
  - **Correlation-based methods:**
    - XGBoost on connectivity features.
    - BrainNetCNN (CNN for brain networks).
    - BrainGNN (graph neural network for brain graphs).
    - BrainNetTF (transformer applied to brain networks).
  - **Voxel-based methods:**
    - TFF: self-supervised transformers for fMRI representation.
    - SwiFT: 4D Swin transformer for fMRI (strong recent voxel-based baseline).
  - All baselines are implemented following their original papers and tuned on the validation sets.

- **Key quantitative findings (trends, not exact numbers):**
  - **Intelligence prediction:**
    - On both HCP and UKB, BrainMT achieves **lower MSE and MAE** and **higher Pearson's R** than all baselines.
    - Many baselines achieve MSE close to 1.0 (given targets are normalized to unit variance), suggesting they mostly predict the mean.
    - BrainMT notably reduces MSE by around **6–9%** relative to the best baselines across datasets, indicating meaningful improvement.
  - **Sex classification:**
    - On HCP, BrainMT achieves the **best accuracy, balanced accuracy, and AUROC** among all methods.
    - On UKB, BrainMT closely matches or slightly exceeds SwiFT, maintaining state-of-the-art performance.
  - **Memory efficiency and scalability:**
    - BrainMT is reported to be about **35.8% more memory-efficient** than SwiFT.
    - Its memory usage grows **linearly** with the number of time frames, enabling longer sequence modeling than standard transformers.
  - **Ablation studies:**
    - **Number of frames:** Using around **200 frames** provides the best trade-off; fewer frames reduce signal-to-noise, while substantially more frames risk overfitting.
    - **Component ablations:** Removing either the transformer or convolution block degrades performance, confirming the importance of the hybrid design.
    - **Depth variations:** Larger or smaller numbers of Mamba/transformer layers change performance; the chosen configuration (12 Mamba, 8 transformer) is near optimal.
    - **Alternative Mamba variants:** Replacing the bidirectional Vision Mamba block with alternatives like VMamba or MambaVision

worsens results, suggesting the specific temporal-first, bidirectional setup is crucial.

        - **Functional connectivity prediction:** BrainMT surpasses SwiFT in subject-level Pearson correlations, indicating better capture of dynamics that underlie functional connectivity.

- **Qualitative / interpretability findings:**
    - **Integrated Gradients (IG) maps** for cognitive intelligence highlight regions in:
        - Default Mode Network (DMN) and Frontoparietal Network (FPN), including posterior cingulate cortex (PCC), anterior cingulate cortex (ACC), precuneus (PCu), and cuneus (Cu).
        - These are known to be involved in working memory, attention, decision-making, and visuospatial processing.
    - For sex prediction, IG maps consistently emphasize regions such as the superior temporal gyrus (STG), middle frontal gyrus (MFG), and precuneus (PCu), aligning with prior studies on sex differences in functional brain organization.
    - These maps suggest that BrainMT's predictive patterns are consistent with established neuroscientific knowledge, not just arbitrary features.

---

8. Strengths, Limitations, and Open Questions

- **Strengths:**
    - **End-to-end voxel-level modeling:** Operates directly on 4D fMRI volumes, avoiding parcellation-induced information loss and inconsistencies.
    - **Long-range temporal modeling:** The Mamba-based temporal-first design enables efficient handling of long sequences (hundreds of frames), which is crucial for capturing slow hemodynamic dynamics.

- **Hybrid architecture:** Combining convolution, Mamba, and transformers gives a balanced treatment of local spatial, long-range temporal, and global spatial dependencies.
- **Strong empirical results:** Consistently outperforms state-of-the-art correlation-based and voxel-based models on large, well-known datasets (HCP and UKB).
- **Interpretability:** Integrated Gradients provide biologically plausible importance maps, grounding model predictions in known functional networks.
- **Memory efficiency:** More memory-efficient than a strong voxel-based transformer baseline (SwiFT), which is important for practical large-scale neuroimaging.

• **Limitations:**

- **Single-modality focus:** The model only uses resting-state fMRI; it does not yet integrate structural MRI, behavioral data, genetics, or other modalities that could aid prediction.
- **Task-specific training:** BrainMT is trained directly on supervised targets rather than via large-scale self-supervised pretraining; this limits its status as a general-purpose foundation model.
- **Compute requirements:** Despite efficiency improvements, training on 4D fMRI with deep Mamba and transformer stacks on large datasets still requires substantial compute (L40S GPUs, distributed training).
- **Generalization beyond studied tasks:** The paper evaluates on sex and cognitive intelligence; it remains unclear how well BrainMT transfers to other phenotypes or clinical conditions without substantial retraining.
- **Interpretability scope:** Integrated Gradients focus on voxel importance but do not fully explain temporal dynamics or causal relationships.

• **Open Questions and Future Directions:**

- **Multimodal extensions:** How would BrainMT perform if extended to jointly model structural MRI, diffusion MRI, or behavioral

measures along with fMRI? Could late fusion or shared embedding approaches from the integration baseline plan improve performance?

- **Self-supervised or foundation-style pretraining:** Can we pretrain BrainMT on large-scale unlabeled rs-fMRI datasets using self-supervised objectives (e.g., masked volume prediction, temporal contrastive tasks) and then fine-tune for many downstream tasks?
- **Clinical translation:** How well does BrainMT generalize to clinical populations (e.g., psychiatric or neurodegenerative disorders), and what adaptations are needed for imbalanced or smaller datasets?
- **Temporal modeling variants:** Would alternative state-space architectures, bidirectionality schemes, or temporal pooling strategies further improve performance or robustness?
- **Uncertainty and reliability:** How can we estimate prediction uncertainty and assess reliability across sites, scanners, and preprocessing pipelines?
- **Integration with graph-based representations:** Could learned voxel-level representations be aggregated into dynamic functional graphs and combined with GNNs for more interpretable connectivity analysis?

---

9. How This Connects to the Bigger Picture (For a New Grad Student)

- **Position in the landscape of brain foundation models:**
  - BrainMT sits in the growing space of **deep backbones for fMRI**, alongside transformer-based models (e.g., TFF, SwiFT) and graph-based methods (BrainGNN).
  - It emphasizes **efficient long-sequence modeling**, borrowing ideas from modern sequence models like Mamba, and adapting them to 4D neuroimaging.

- While not a traditional “foundation model” with massive pretraining, it can serve as a **strong architectural template** for future brain FMs.

- **Connections to well-known ideas:**
  - Conceptually, BrainMT is like **“a video model for brain volumes”**: it treats fMRI as a spatiotemporal sequence, similar to video transformers but with neurobiological constraints.
  - The Mamba block provides **linear-time sequence modeling**, akin to efficient alternatives to transformers in NLP and vision; the transformer layer then adds global relational reasoning.
  - Compared to traditional connectivity pipelines, it replaces hand-designed ROI-based features with **end-to-end learned representations** from raw data.
- **Relation to integration and broader research programs:**
  - For multimodal integration projects (e.g., integrating brain imaging with genomics or clinical data), BrainMT can act as a **strong fMRI encoder**, producing subject-level representations that can be fused with other modalities via late fusion or joint embedding methods.
  - Its success supports the integration baseline plan’s principle of **preserving modality-specific signal** (learning rich fMRI representations before combining with other data types).
  - The evaluation practices (consistent splits, multiple metrics, ablations) echo the plan’s emphasis on careful, robust benchmarking.
- **Why this paper is a useful reference:**
  - It provides a concrete, well-validated example of how to design and train a modern deep architecture for 4D fMRI at scale.
  - It offers design patterns (temporal-first scanning, hybrid Mamba–transformer stacks, integrated interpretability) that can inform future brain FMs and multimodal models.

    - For a new grad student, it is a good entry point into understanding how sequence modeling ideas from NLP/vision can be translated into neuroimaging.

---

10. Key Takeaways (Bullet Summary)

- **Problem:** Resting-state fMRI contains complex, long-range spatiotemporal patterns that can predict subject-level phenotypes such as sex and cognitive intelligence, but existing approaches either lose spatial detail (via parcellation) or are limited to short time windows (due to transformer complexity).
- **Problem:** There is a need for architectures that can handle long 4D fMRI sequences efficiently while capturing both local and global dependencies in brain activity.
- **Method / model:** BrainMT is a **hybrid deep architecture** combining convolutional blocks, a bidirectional Vision Mamba state-space module with temporal-first scanning, and a transformer block for global spatial attention.
- **Method / model:** The model processes 200 fMRI frames per subject at a reduced spatial resolution, builds a long token sequence with positional embeddings and a classification token, and uses Mamba to capture long-range temporal context before applying transformer attention.
- **Method / model:** Training uses large-scale resting-state fMRI datasets (UKB and HCP) with supervision for sex classification and cognitive intelligence regression, optimized with AdamW and standard loss functions.

- **Results:** BrainMT outperforms strong correlation-based (XGBoost, BrainNetCNN, BrainGNN, BrainNetTF) and voxel-based (TFF, SwiFT) baselines in both intelligence prediction and sex classification across HCP and UKB.
- **Results:** The model is more memory-efficient than SwiFT, scales linearly with the number of frames, and benefits from longer temporal context; ablations confirm the importance of its hybrid design and the specific Mamba variant.
- **Results:** Integrated Gradients reveal that BrainMT's predictions rely on default mode and frontoparietal regions for intelligence, and regions like STG, MFG, and precuneus for sex differences, aligning with known neuroscience.
- **Why it matters:** BrainMT demonstrates that modern sequence models like Mamba, combined with transformers, can efficiently model long 4D fMRI sequences and yield state-of-the-art predictive performance.
- **Why it matters:** The architecture is a promising backbone for future brain foundation models and a strong fMRI encoder that could be integrated into larger multimodal systems (e.g., combining brain imaging with genetics or clinical data).