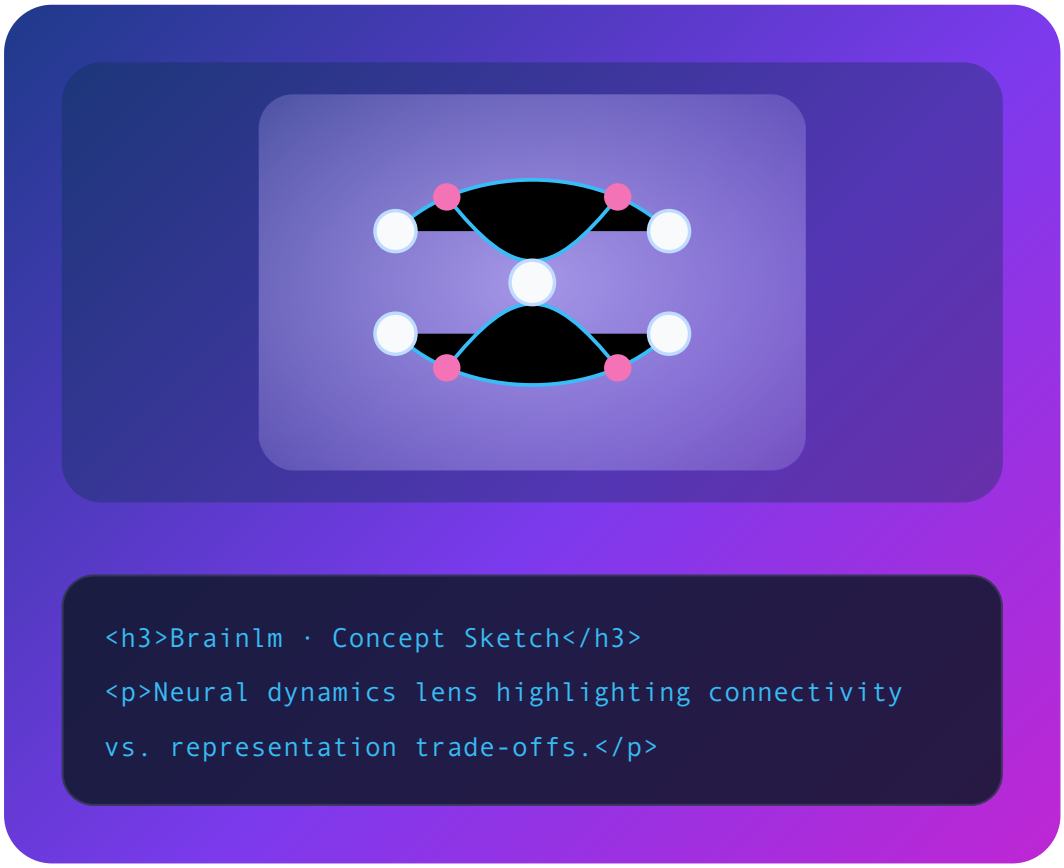

BrainLM: A Foundation Model For Brain Activity Recordings

Authors: Josue Ortega Caro, Antonio H. de O. Fonseca, Syed A. Rizvi, Matteo Rosati, Christopher Averill, James L. Cross, Prateek Mittal, Emanuele Zappala, Rahul M. Dhodapkar, Chadi G. Abdallah, David van Dijk, et al.
Year: 2024
Venue: ICLR (International Conference on Learning Representations)

1. Classification

- **Domain Category:**
  - **Brain FM.** The paper develops a large foundation model specifically for functional MRI (fMRI) recordings, learning

spatiotemporal representations of brain activity dynamics across the whole brain.

- **FM Usage Type:**
  - **Core FM development.** BrainLM is introduced as a new foundation model architecture and pretraining scheme for fMRI, with subsequent fine-tuning and zero-shot applications.
- **Key Modalities:**
  - Task-based and resting-state **fMRI** (BOLD time series) from large population cohorts (UK Biobank and Human Connectome Project).

---

2. Executive Summary

This paper introduces **BrainLM**, a large transformer-based foundation model trained on **6,700 hours of fMRI recordings** from over 77,000 scans. Instead of training separate models for each narrow decoding task, BrainLM is pretrained in a **self-supervised masked-reconstruction** fashion to learn general-purpose representations of whole-brain activity over time. After pretraining, the same model can be fine-tuned to predict **clinical variables** (age, neuroticism, PTSD, anxiety), **forecast future brain states**, and perform **zero-shot inference** such as discovering functional brain networks directly from attention patterns. The authors show that BrainLM **generalizes across datasets**, performing well both on held-out UK Biobank scans and on the independent Human Connectome Project cohort. They also demonstrate interpretable attention maps that align with known brain networks and clinical differences (e.g., depression severity). For a new grad student, this paper is a key example of how **foundation model ideas from language and vision** (e.g., masked autoencoders) can be adapted to **neuroimaging**, creating a versatile model that unifies many downstream tasks on fMRI data.

---

3. Problem Setup and Motivation

- **Scientific / practical problem**
  - Build a **single, general-purpose model** of brain activity dynamics that can support many downstream tasks: predicting clinical variables, modeling brain networks, and forecasting future activity.
  - Learn **unsupervised representations** from large-scale fMRI repositories rather than training separate, task-specific models on small datasets.
  - Capture the **full spatiotemporal structure** of fMRI signals across the brain, not just limited regions like the visual cortex.
- **Why this is hard**
  - **High dimensionality and complexity:** fMRI produces thousands of voxel or parcel time series, with complex dependencies across brain regions and time.
  - **Indirect and noisy signal:** BOLD signals are an indirect measure of neural activity and can be hard to interpret.
  - **Limited labels:** Many large fMRI datasets have rich time series but relatively few labels for specific tasks, making supervised training challenging.
  - **Task-specific models do not generalize well:** Traditional supervised models (e.g., SVMs, small neural nets) are tuned to narrow tasks and do not transfer well to new datasets or objectives.
  - **Need for scalable training:** To benefit from large repositories like UK Biobank and HCP, models must handle massive data and learn representations that scale with model size and data size.

4. Data and Modalities

- **Datasets used**
  - **UK Biobank (UKB):**
    - ~**76,296** task-based and resting-state fMRI recordings with associated medical records.
    - Ages approximately 40–69; scanned on a Siemens 3T scanner at ~0.735 s temporal resolution.
    - **80% (61,038 recordings)** used for training; 20% held out for testing.
  - **Human Connectome Project (HCP):**
    - **1,002** high-quality fMRI recordings from healthy adults.
    - ~0.72 s temporal resolution; used entirely as an external evaluation cohort.
  - In total, the training corpus spans **77,298 recordings** and **6,700 hours** of preprocessed fMRI.
- **Modalities**
  - Single modality: **functional MRI (fMRI)**, representing whole-brain BOLD time series.
  - Both **task** and **resting-state** recordings are included.
- **Preprocessing / representation**
  - Standard preprocessing: **motion correction**, **normalization**, **temporal filtering**, and **ICA-based denoising**.
  - Brain parcellation into **424 regions (AAL-424 atlas)**, yielding 424-dimensional time series per scan.
  - Time series sampled at ~1 Hz after preprocessing.
  - **Robust scaling**: per-parcel median subtraction and division by interquartile range across subjects.
  - For model input:
    - Random **200-timestep subsequences** are extracted from each recording.

        - Each parcel's 200-timestep sequence is split into **patches of 20 time points**, giving 10 patches per parcel.
        - The resulting patches (conceptually 424 × 10) are treated as "tokens" via a **learnable linear projection** into 512-dimensional embeddings.
        - The 424 × 200 window is also viewed as a **2D image** (parcels × time) with parcels ordered by Y-coordinate to preserve spatial locality.
- **Missing details**
    - Exact number of subjects, hardware details beyond scanner type, and some hyperparameters are referenced but not fully spelled out in the main extracted text (likely given in supplementary material).

---

5. Model / Foundation Model

- **Model Type**
    - **Masked autoencoder (MAE) based on a Transformer** architecture.
    - Encoder–decoder structure with **self-attention blocks** that operate on spatiotemporal tokens derived from parcel-time patches.
- **Is it a new FM or an existing one?**
    - **New foundation model.** The authors design BrainLM specifically for fMRI data, inspired by **BERT** and **Vision Transformer (ViT)** style masked modeling but adapted to 2D parcel×time structure.
- **Key components and innovations**

Component	Description
Tokenization of fMRI patches	200-timestep windows split into 20-timestep patches per parcel; patches projected into 512-d embeddings.
Spatiotemporal masking	Random and future-timepoint masking at rates of **20%, 75%, or 90%**, making the model reconstruct masked tokens.
2D “image-like” formulation	Treats the 424-parcel × 200-timestep window as a 2D grid; parcels ordered by Y-coordinate to preserve spatial adjacency, enabling multi-parcel tokens and scalable encoding.
Transformer encoder	Processes **only unmasked tokens**, with **4 self-attention layers** and **4 attention heads**.
Transformer decoder	**2-layer decoder** that takes both encoded visible tokens and masked tokens, then reconstructs the full input.
Positional embeddings	Learnable **spatial and temporal embeddings** added to token representations to encode parcel location and time.
Latent CLS token	A special token summarizing each sequence, later used for clinical variable prediction and visualization.

- **Training setup**
  - **Objective:** Minimize **mean squared error (MSE)** between original and reconstructed fMRI signals for masked patches (self-supervised reconstruction).
  - **Pretraining data:** 6,700 hours of fMRI from UKB and HCP, using random 200-timestep subsequences.
  - **Optimization: Adam** optimizer; **100 epochs**; batch size **512**.
  - **Scaling:** Multiple model sizes (e.g., **13M, 111M, 650M parameters**), with performance improving as both model size and dataset size increase.
  - **Downstream adaptation:**
    - Add a **3-layer MLP head** to the pretrained encoder for regression of clinical variables.
    - Fine-tune on subsets of UKB data withheld from pretraining.
    - For future state prediction, fine-tune the model to forecast the next 20 timesteps given 180 observed timesteps.

---

6. Multimodal / Integration Aspects (If Applicable)

- **Is the paper multimodal?**
  - **Not in the main experiments.** BrainLM is trained and evaluated on **single-modality fMRI data** (task and rest). The core contribution is a **unimodal** foundation model for brain activity recordings.
  - However, the **discussion explicitly points to multimodal extensions** as future work, suggesting integration with EEG, MEG, and other brain-wise or even genomic information.

- **Relation to integration baseline plan**
  - The paper itself does **not** implement late fusion, CCA, or contrastive cross-modal alignment. It is focused on learning a strong **single-modality encoder** for fMRI.
  - In terms of the integration baseline plan:
    - BrainLM can be seen as a **per-modality encoder** that could feed into a **late fusion** or **stacking** approach when combined with other modalities (e.g., structural MRI, genetics, clinical variables).
    - Its robust self-supervised representations align with the plan's emphasis on **preserving modality-specific signal** before fusion.
    - The evaluation on multiple tasks and datasets is compatible with the plan's emphasis on **robustness and disciplined evaluation**, although the paper does not explicitly follow the full AUROC/AUPRC + confidence-interval protocol discussed in the plan.
  - Future multimodal systems could:
    - Use BrainLM's embeddings as one tower in a **two-tower contrastive model**, with another tower encoding, for example, genetics or behavioral data.
    - Perform **late fusion** of BrainLM features with those from other FMs (e.g., for combined clinical prediction).
- **Summary**
  - For now, BrainLM is best viewed as a **strong building block for multimodal integration**, rather than a multimodal FM itself.

7. Experiments and Results

- **Tasks / benchmarks**
  - **Masked reconstruction / generalization:**
    - Evaluate reconstruction accuracy (e.g., R^2 on masked patches) on held-out **UKB test data** and independent **HCP** data.
  - **Clinical variable prediction:**
    - Fine-tune BrainLM to regress **age**, **neuroticism**, **PTSD (PCL-5)**, and **general anxiety (GAD-7)** scores from fMRI recordings.
  - **Future brain state prediction:**
    - Given 180 observed timesteps, predict the next 20 timesteps of parcel activity, evaluated on UKB and HCP.
  - **Interpretability via attention analysis:**
    - Analyze self-attention weights (especially from the CLS token) to study how attention changes across tasks and clinical groups (e.g., depression severity).
  - **Functional network prediction (zero-shot-like):**
    - Use attention-derived features to classify parcels into **7 intrinsic functional networks** without network-specific supervision.
- **Baselines**
  - For clinical variable regression:
    - **SVR** and **MLP** on correlation matrices.
    - **LSTM** and **GCN** models that directly use fMRI recordings.
    - Comparisons both to models trained on **raw data** and models on **pretrained embeddings**.
  - For future state prediction:
    - **LSTM**.
    - **Neural ODE** and **Latent ODE** models.
    - A **Transformer model without pretraining** (same architecture but trained only on the forecasting task).

    - For functional network identification:
        - k-NN classifiers using:
            - Raw parcel time series,
            - Variational Autoencoder (VAE) embeddings,
            - GCN embeddings,
            - BrainLM attention weights.
- **Key findings**
    - **Generalization and reconstruction:**
        - BrainLM achieves strong R^2 on UKB test data and **generalizes well to HCP**, despite domain differences, showing that pretraining learns robust, dataset-agnostic representations.
        - Performance improves with **larger models and more data**, demonstrating **scaling laws** similar to those seen in language and vision FMs.
    - **Clinical variable prediction:**
        - BrainLM-based regressors achieve **lower mean squared error** than baselines (SVR, MLP, LSTM, GCN, raw data) across age, PTSD, anxiety, and neuroticism.
        - Fine-tuning further improves over using frozen embeddings, indicating that pretrained representations are **rich but still adaptable**.
        - Even **zero-shot regression** (no fine-tuning) shows non-trivial predictive power, and performance scales with model size.
    - **Future brain state prediction:**
        - Fine-tuned BrainLM **significantly outperforms LSTM, Neural ODE, Latent ODE, and the non-pretrained Transformer** on both UKB and HCP.
        - The benefit of pretraining is clear: the model without pretraining performs noticeably worse.
        - Larger BrainLM variants maintain better forecasting performance over multiple timesteps.

- **Interpretability and networks:**
  - Attention maps distinguish **task vs rest** (e.g., stronger attention to visual cortex during task states).
  - Differences in attention for **high vs low depression** emphasize frontal and limbic regions, consistent with clinical literature.
  - Using attention-based features, BrainLM achieves **~58.8% accuracy** in classifying parcels into 7 functional networks, outperforming VAE, GCN, and raw data baselines.

---

8. Strengths, Limitations, and Open Questions

- **Strengths**
  - Introduces a **true foundation model for fMRI**, trained at a scale (6,700 hours, 77k recordings) much larger than prior work.
  - Uses **self-supervised masked modeling** to efficiently exploit unlabeled fMRI data, enabling versatile downstream applications.
  - Demonstrates **strong generalization** across cohorts (UKB → HCP) and across diverse tasks (reconstruction, forecasting, clinical prediction, network identification).
  - Provides **interpretable attention maps** that align with known brain networks and clinical patterns, offering neuroscientific insights.
  - Shows clear **scaling behavior**, suggesting that larger models and datasets can further improve performance.
- **Limitations**
  - Currently **unimodal**: only fMRI is modeled; multimodal integration with EEG, structural MRI, genetics, and behavior is left for future work.
  - Despite interpretability via attention, the latent representations are still **complex**, and a full mechanistic understanding of what is encoded remains challenging.

- The approach is **computationally heavy**, requiring large-scale pretraining with transformers on big imaging datasets.
- Some preprocessing choices (e.g., parcellation scheme, scaling, window length) may influence results, but not all ablations are detailed in the main text.
- Real-world clinical deployment requires careful validation, robustness checks, and fairness analyses that go beyond the current experiments.

- **Open Questions and Future Directions:**
  - How does BrainLM compare to **alternative pretraining objectives** (contrastive, masked prediction on different views, generative modeling) for fMRI?
  - Can BrainLM embeddings be effectively **combined with other modalities** (EEG, MEG, structural MRI, genetics) using late fusion or contrastive two-tower setups, and does this improve clinical prediction?
  - What **neuroscientific structure** is captured in the CLS token and internal layers—can we relate specific attention patterns or latent dimensions to known circuits or cognitive processes?
  - How robust is BrainLM to **distribution shifts** such as different scanners, acquisition protocols, or clinical populations (e.g., pediatric, elderly, or specific disorders)?
  - Can smaller, **distilled versions** of BrainLM retain most performance while being practical for clinical or real-time applications?

9. Context and Broader Impact

- **Position in the FM landscape**
  - BrainLM is to **fMRI** what large language models (like GPT) are to text: a **general, pretrained backbone** that can be adapted to many tasks rather than a task-specific model.
  - Within brain/neuro FMs, it extends prior work that focused on **visual cortex or small datasets** to **whole-brain modeling at population scale**.
  - It shows that the **masked autoencoder paradigm** from vision and language transfers well to **spatiotemporal brain data**.
- **Relation to well-known ideas**
  - Conceptually, BrainLM behaves like a **BERT-style or ViT-style masked model** applied to a 2D grid where one dimension is space (brain parcels) and the other is time (fMRI timesteps).
  - The attention-based interpretability parallels how we interpret attention maps in NLP and vision, but here the “tokens” are **brain parcels over time**.
  - The clinical prediction and network discovery tasks illustrate how **foundation models can support both prediction and scientific discovery**.
- **Why it matters and how it links to integration plans**
  - For a grad student interested in **computational neuroscience**, BrainLM is a blueprint for building large, reusable models of brain activity.
  - It provides a **ready-made encoder** that can plug into multimodal integration pipelines—consistent with the integration baseline plan’s idea of learning strong, modality-specific representations before fusion.
  - The work signals a general trend: **foundation models are moving into neuroimaging**, opening paths to richer multimodal systems that combine brain signals with genetic, behavioral, and clinical data.

10. Key Takeaways (Bullet Summary)

- **Problem**
  - There is a need for a **single, scalable model** that can learn from massive fMRI repositories and support many downstream neuroscience and clinical tasks.
  - Traditional task-specific models struggle with **generalization, data scale, and transfer across cohorts**.
- **Method / model**
  - BrainLM is a **transformer-based masked autoencoder** that treats fMRI parcel×time windows as a 2D grid of tokens.
  - It uses **spatiotemporal masking and reconstruction** to learn representations from 6,700 hours of fMRI without task labels.
  - The architecture includes **4-layer encoder and 2-layer decoder** transformers, with learned spatial and temporal embeddings and a summary **CLS token**.
  - Multiple model sizes (13M–650M parameters) are trained, showing **improved performance with scale**.
- **Results**
  - BrainLM shows strong **reconstruction and generalization** performance on both UKB and HCP datasets.
  - It **outperforms baselines** (SVR, MLP, LSTM, GCN, Neural ODE, non-pretrained Transformer) in clinical variable prediction and future brain state forecasting.
  - Attention-based analyses reveal **meaningful functional networks and clinical differences**, and attention-derived features outperform other representations in classifying parcels into known networks.

- **Why it matters**
  - BrainLM establishes a **foundation model paradigm for fMRI**, demonstrating that large, self-supervised models can unify diverse tasks in brain dynamics modeling.
  - It provides a **flexible, interpretable, and extensible backbone** that future work can extend to multimodal settings and more ambitious clinical and neuroscientific applications.

Generated via custom pipeline · 2025-11-20