bioRxiv preprint doi: https://doi.org/10.1101/2023.09.12.557460; this version posted September 13, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

BrainLM: A foundation model for brain activity recordings

Josue Ortega Caro *[1], Antonio H. de O. Fonseca * [2], Christopher Averill [3], Syed A. Rizvi [4], Matteo Rosati [5], James L. Cross [6], Prateek Mittal [7], Emanuele Zappala [12], Daniel Levine [4], Rahul M. Dhodapkar [8], Chadi G. Abdallah [3,**], and David van Dijk [1,2,4,9,10,11,**]

[1]Wu Tsai Institute, Yale University
[2]Interdepartmental Neuroscience Program, Yale University
[3]Baylor College of Medicine
[4]Department of Computer Science, Yale University
[5]Yale School of Medicine
[6]Yale University
[7]Thapar Institute of Engineering and Technology
[8]University of Southern California
[9]Interdepartmental Program in Computational Biology & Bioinformatics, Yale University
[10]Internal Medicine, Yale University
[11]Cardiovascular Research Center, Yale University
[12]Department of Mathematics and Statistics, Idaho State University
[**]Co-last Author, Corresponding Authors: chadi.abdallah@bcm.edu, david.vandijk@yale.edu

Abstract

We introduce the Brain Language Model (BrainLM), a foundation model for brain activity dynamics trained on 6,700 hours of fMRI recordings. Utilizing self-supervised masked-prediction training, BrainLM demonstrates proficiency in both fine-tuning and zero-shot inference tasks. Fine-tuning allows for the prediction of clinical variables and future brain states. In zero-shot inference, the model identifies functional networks and generates interpretable latent representations of neural activity. Furthermore, we introduce a novel prompting technique, allowing BrainLM to function as an *in silico* simulator of brain activity responses to perturbations. BrainLM offers a novel framework for the analysis and understanding of large-scale brain activity data, serving as a "lens" through which new data can be more effectively interpreted.

1 Introduction

Understanding how cognition and behavior arise from brain activity stands as one of the fundamental challenges in neuroscience research today. Functional magnetic resonance imaging (fMRI) has emerged as a critical tool for pursuing this goal by providing a noninvasive window into the working brain. fMRI measures blood oxygen level fluctuations that reflect regional neural activity. However, analyzing the massive, high-dimensional recordings produced by fMRI poses major challenges. The blood-oxygen-level dependent (BOLD) signals represent an indirect measure of brain function and can be difficult to interpret. Furthermore, fMRI data exhibits complex spatiotemporal dynamics, with critical dependencies across both space and time. Most existing analysis approaches fail to fully model these complex nonlinear interactions within and across recordings [1].

Prior fMRI analysis techniques have relied heavily on machine learning models designed for specific narrow tasks [2, 3, 4], hindering their generalizability. Traditional models also struggle to integrate information across the wealth of unlabeled fMRI data available. Hence, there is an ongoing need for flexible modeling frameworks that can truly capitalize on the scale and complexity of fMRI repositories.

[*] Equal Contribution

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.12.557460; this version posted September 13, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

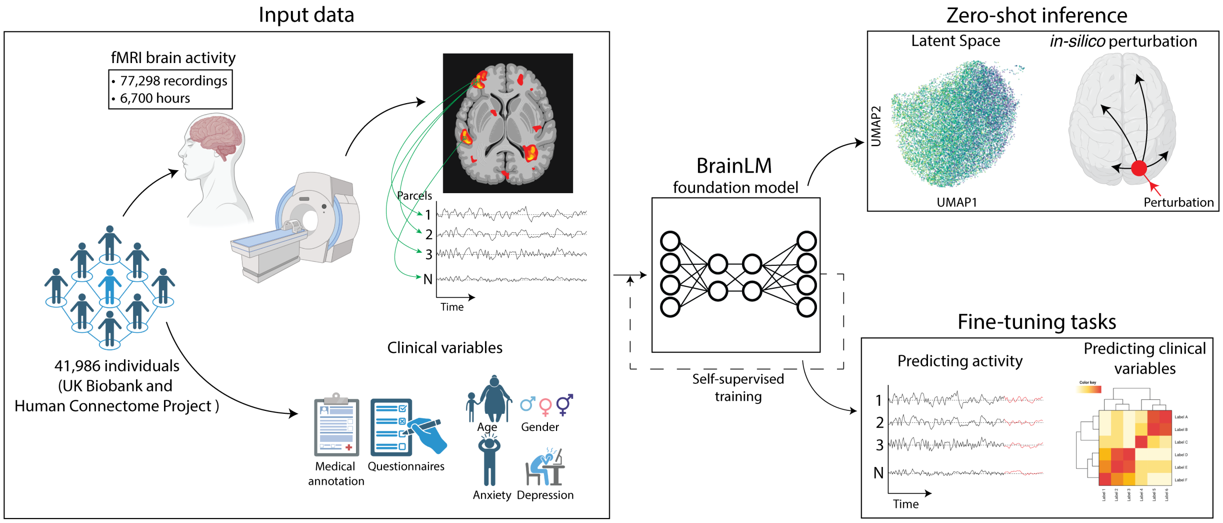

Figure 1: Overview of the BrainLM framework. The model was pretrained on 6,700 hours of fMRI recordings from 77,298 subjects via spatiotemporal masking and reconstruction. After pretraining, BrainLM supports diverse capabilities through fine-tuning and zero-shot inference. Fine-tuning tasks demonstrate prediction of future brain states and clinical variables from recordings. Zero-shot applications include inferring functional brain networks from attention weights and using a novel prompting technique to simulate perturbation responses in silico. This highlights BrainLM's versatility as a foundation model for fMRI analysis.

Foundation models represent a new paradigm in artificial intelligence, shifting from narrow, task-specific training to more general and adaptable models [5]. Inspired by breakthroughs in natural language processing, the foundation model approach trains versatile models on broad data at scale, enabling a wide range of downstream capabilities via transfer learning. Unlike previous AI systems designed for singular functions, foundation models exhibit general computational abilities that make them suitable for myriad real-world applications. Large language models like GPT have demonstrated the potential of this framework across diverse domains including healthcare, education, robotics, and more [6, 7, 8, 9]. Foundation models offer new opportunities to rethink challenges in neuroscience and medical imaging analysis.

Here, we introduce BrainLM, the first foundation model for fMRI recordings. BrainLM leverages a Transformer-based architecture to capture the spatiotemporal dynamics inherent in large-scale brain activity data. Pretraining on a massive corpus of raw fMRI recordings enables unsupervised representation learning without task-specific constraints (see Figure 1).

After pretraining, BrainLM supports diverse downstream applications via fine-tuning and zero-shot inference. We demonstrate BrainLM's capabilities on key tasks including prediction of future brain states, decoding cognitive variables, discovery of functional networks, and in silico perturbation analysis. Together, these results highlight BrainLM's proficiency in both zero-shot and fine-tuning tasks. This work highlights the potential of applying large language models to advance neuroscience research. BrainLM is a foundation model for the community to build upon, providing more powerful computational tools to elucidate the intricate workings of the human brain.

2 Related Work

Prior work has explored various machine-learning techniques for analyzing fMRI recordings. Earlier approaches focused on decoding cognitive states from activity patterns. Methods like SVM and neural networks were trained in a supervised fashion to classify fMRI data into stimulus categories or regress against variables of interest [10, 11, 12]. However, these models learn representations tailored to specific tasks and struggle to generalize.

Recent work has aimed to obtain more transferable fMRI encodings without task-specific constraints. Techniques include training autoencoders to reconstruct recordings, learning to map recordings to a lower-dimensional space [2, 3, 4]. However, most methods operate on small datasets, limiting their ability to learn robust generalizable representations.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.12.557460; this version posted September 13, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Our work is most closely related to recent efforts to apply masked autoencoders for unsupervised pretraining on fMRI data [13, 14] or other neural recording modalities [15, 16, 17]. However, these prior MAE models are trained on 1-2 orders of magnitude less data compared to BrainLM. The scale of our dataset combined with a Transformer-based architecture enables BrainLM to learn substantially more powerful encodings of spatiotemporal fMRI patterns.

With relation to the *in silico* experiments, our approach follows in similarity to other methods applied to stimulus optimization [18, 19, 20, 21], but in our case, it is for optimizing neural dynamics.

Critically, BrainLM is the first model of its scale designed following the foundation model paradigm - pretrained on diverse unlabeled data and adaptable to various downstream applications via transfer learning. This distinguishes BrainLM from prior work and demonstrates the potential of large foundation models for fMRI analysis. In summary, our innovations include training a masked autoencoder at an unprecedented scale, employing a Transformer architecture suited for spatiotemporal data, and demonstrating versatile fine-tuning and interpretation applications enabled by the foundation model approach.

3 Methods

3.1 Datasets and Preprocessing

We leveraged two large-scale publicly available datasets - the UK Biobank (UKB) [22] and the Human Connectome Project (HCP) [23]. The UKB provides task-based and resting-state functional MRI (fMRI) recordings plus medical records from over 40,000 subjects aged 40-69 years old. recordings were acquired on a Siemens 3T scanner at 0.735s temporal resolution. The HCP contains 1,002 high-quality fMRI recordings from healthy adults scanned at 0.72s resolution.

Our model was trained on 80% of the UKB dataset (61,000 recordings) and evaluated on the held-out 20% and the full HCP dataset. All recordings underwent standard preprocessing including motion correction, normalization, temporal filtering, and ICA denoising to prepare the data [24, 25].

To extract parcel-wise time series, we parcellated the brain into 424 regions using the AAL-424 atlas [26]. This yielded 424-dimensional scan sequences sampled at 1 Hz. Robust scaling was applied by subtracting the median and dividing by the interquartile range computed across subjects for each parcel.

In total, our training dataset comprised 6,700 hours of preprocessed fMRI activity patterns from 77,298 recordings across the two repositories. This large-scale corpus enabled unsupervised pretraining of BrainLM to learn robust functional representations.

3.2 Model Architecture

Our model consists of a Transformer-based [27] masked autoencoder architecture adapted from natural language processing models like BERT [28] and Vision Transformer [29]. The encoder comprises a stack of multi-headed self-attention layers that operate on visible (unmasked) input patches. The decoder reconstructs the original patches, including masked ones, by attending to the encoder activations.

During training, segments of parcel time series are embedded and then partially masked to represent missing data. The unmasked segments are encoded and decoded to reconstruct the full input. By training to minimize the reconstruction error of masked segments, the model learns dependencies within and across brain activity sequences (see Figure 2).

Implementation details can be found in the supplement. The code, model weights and training hyperparameters will be made publicly available.

3.3 Training Procedure

For each fMRI recording, we sampled random 200-timestep subsequences. The parcel time series were divided into segments of 20 timesteps, yielding 10 segments per subsequence. These were embedded into a 512-dimensional space and masked with a ratio of 20%, 50%, or 75%.

The unmasked segments were encoded via a Transformer encoder with 4 self-attention layers and 4 heads. This was decoded by a 2-layer Transformer to reconstruct all segments. We trained with batch size 512 and the Adam optimizer for 100 epochs, minimizing the mean squared error between original and predicted embeddings (see

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.12.557460; this version posted September 13, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 2). After pretraining on all sequences, the encoder can extract informative features capturing spatiotemporal brain activity patterns. We leverage the pre-trained encoder for downstream prediction and interpretation tasks.

3.4 Clinical Variable Prediction

As one downstream application, we fine-tuned BrainLM to predict clinical variables from fMRI recordings. The pretrained encoder was appended with a 3-layer MLP head and trained to regress targets including age, neuroticism, PTSD, and anxiety disorder scores. For age, we normalize by simply Z-scoring the Age values for all patients to a mean of 0 and unit variance. For Neuroticism scores, we do min-max scaling to bring the distribution of scores into the range $[0,1]$. For Post Traumatic Stress Disorder (PCL) and General Anxiety Disorder (GAD7) scores, we first perform a log transformation to make the values less exponentially distributed, and then perform mix-max scaling to range.

We perform clinical variable regression on data unseen by BrainLM during fine-tuning. We use a split of held-out samples from the UK Biobank test set as the dataset for fine-tuning BrainLM and training SVM regressors, and we evaluate the performance against baseline models using raw input data and extracted pretrained embeddings. During fine-tuning, we apply a 40% dropout on the activations of both the BrainLM encoder and MLP head to prevent overfitting.

3.5 *In Silico* Perturbation Simulation

A key element of current foundation models is prompting. We introduce a novel in silico perturbation prompting approach to elucidate the model's representation of brain dynamics. This technique simulates the effect of perturbations on neural activity patterns in a completely computational and "zero-shot" manner, i.e. the model was never trained on perturbations.

We define a perturbation function $G(X_{input}, \theta)$ with parameters θ that modifies the input fMRI sequence $(\tilde{X})$. The perturbed sequence is fed into the pre-trained BrainLM and θ optimized to minimize the distance between the predicted state $(\hat{X})$ and a target (X_T) corresponding to a cognitive condition of interest (see Figure 6 and Algorithm 1). By analyzing the optimized perturbations $(\tilde{X} - X_{input})$, we can identify which fMRI features are most influential in altering the model's predicted brain state. This reveals insights into relationships learned by BrainLM.

We applied this *in silico* simulation to probe differences between resting and task states. The optimized perturbations highlighted BrainLM's sensitivity to task-driven changes in visual cortical areas.

4 Results

4.1 Model Generalization

To evaluate BrainLM's ability to generalize to new fMRI data, we tested its performance on held-out recordings from both the UKB test set as well as the independent HCP dataset.

On the UKB test data, BrainLM achieved an average R^2 score of 0.402, indicating strong generalization on unseen recordings from the same distribution. We also found that key training hyperparameters impacted generalization - larger training data and 75% masking ratio yielded the best performance on the test set (see Table 1).

Critically, BrainLM also generalized well to the HCP dataset, achieving a similar R^2 score of 0.316. Despite differences in cohort and acquisition details, BrainLM could effectively model brain dynamics in this entirely new distribution. Figure 3 shows sample reconstructions on both UKB and HCP recordings. Overall, these results demonstrate BrainLM's ability to learn robust representations of fMRI recordings that generalize across datasets. This highlights the advantages of large-scale pretraining for learning widely applicable models of brain activity.

4.2 Prediction of Clinical Variables

A key advantage of foundation models is their ability to fine-tune on downstream tasks using the pretrained representations. We evaluated BrainLM's clinical prediction capabilities by fine-tuning to regress metadata variables obtained from the UKBioBank dataset.

The pretrained encoder was appended with an MLP head and fine-tuned to predict age, neuroticism, PTSD, and anxiety disorder scores. fine-tuning used a held-out subset of UKB subjects. BrainLM's performance was compared against: 1) SVMs trained on raw fMRI data, and 2) SVMs trained on BrainLM's pretrained embeddings. Across all variables, the fine-tuned BrainLM model achieved significantly lower mean squared error compared

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.12.557460; this version posted September 13, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

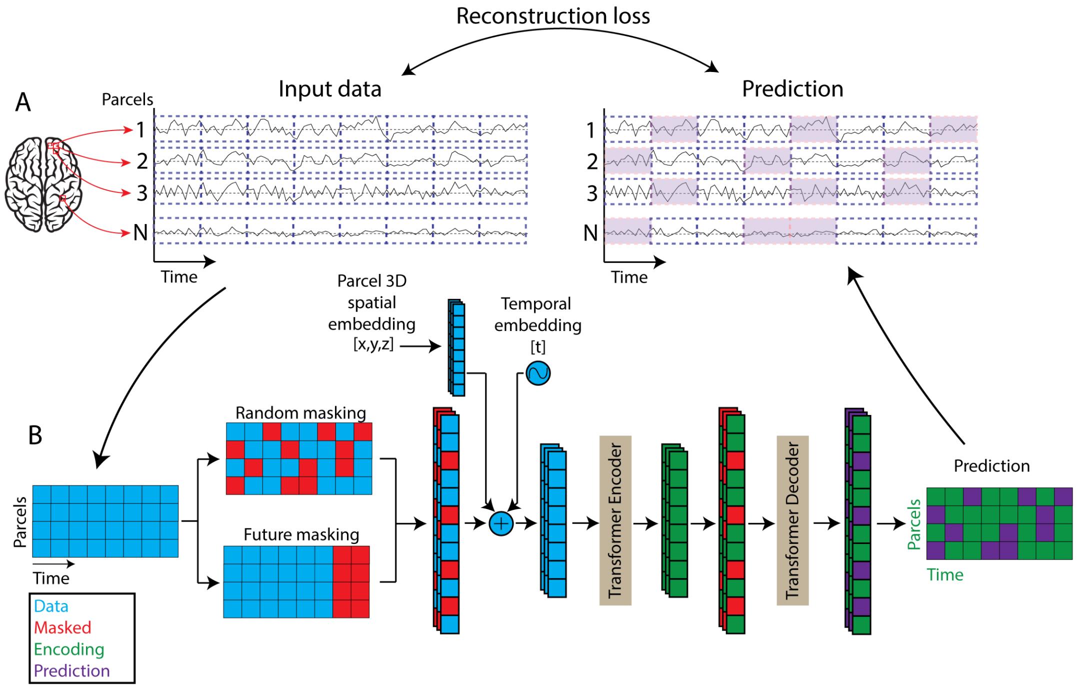

Figure 2: BrainLM architecture and training procedure. A) The fMRI recordings are compressed into 424 dimensions (parcels) (See Methods 3.1). The recordings are randomly trimmed to 200 time points. For each parcel, the temporal signal is split into patches of 20 time points each (blue dashed boxes). The resulting 4240 patches are converted into tokens via a learnable linear projection. B) From the total number of tokens (blue), a subset is masked (red), either randomly or at future timepoints. We then add the learnable spatial and temporal embeddings to each token. These visible tokens (blue) are then processed by a series of Transformer blocks (Encoder). The input to the Decoder is the full set of tokens, consisting of encoded visible tokens (green) and masked tokens (red). The Decoder also consists of Transformer blocks and ultimately projects the tokens back to data space. Finally, we compute the reconstruction loss between the prediction (purple) and the original input data (blue).

Table 1: Performance comparison of latent space learned through self-supervised pretraining. Shown is the coefficient of determination (R^2) between predicted and ground truth data for masked patches across various configurations of masking ratio (MR) and training data size. Columns indicate models trained on 1% or 100% of the data and with 75% or 90% masking. Rows show different inference metrics, validated by masking 20%, 50%, or 75% on UK Biobank data (UKB) or Human Connectome Project (HCP) data. The top performing model was trained on 100% of the data with 75% making.

	MR=0.75		MR=0.90	
Data size	1%	100%	1%	100%
UKB (MR=0.2)	0.361	**0.402**	0.158	0.221
UKB (MR=0.5)	0.349	**0.389**	0.182	0.245
UKB (MR=0.75)	0.309	**0.343**	0.186	0.234
HCP (MR=0.2)	0.300	**0.316**	0.126	0.176

to the baselines (see Table 2). Notably, even the pretrained model outperformed the raw data SVM, indicating BrainLM's representations better capture informative clinical biomarkers. Additional gains from fine-tuning further demonstrate the benefits of initializing with meaningful pretrained features before tuning to a specific prediction task.

Overall, these results validate BrainLM's ability to uncover predictive signals within complex fMRI recordings. By leveraging large-scale pretraining and transfer learning, BrainLM offers a powerful framework for fMRI-based

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.12.557460; this version posted September 13, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

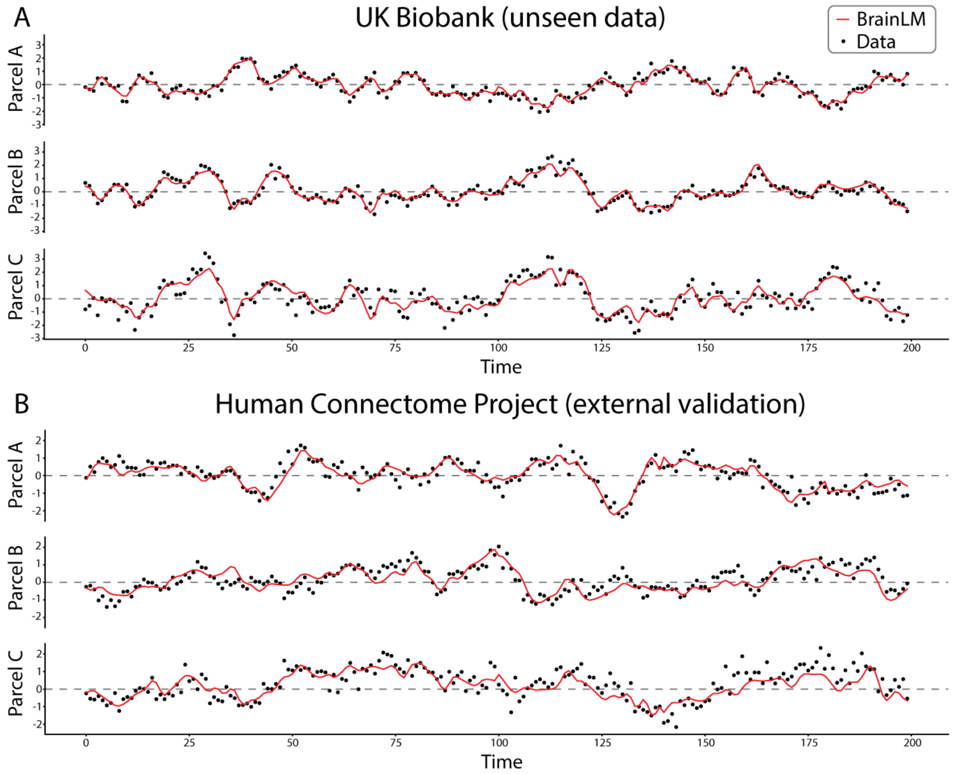

Figure 3: BrainLM reconstruction performance on held-out data. The model predictions (red) closely fit the ground truth recordings (black) of unseen data sampled from the cohort that the model was trained on (A, UKB) as well as data sampled from an external never-before-seen cohort (B, HCP). This demonstrates BrainLM's ability to generalize across subjects and datasets.

Table 2: Results for the regression of the clinical features. The values show the MSE (mean ± std)

	Age	Post traumatic stress disorder (PCL)	General Anxiety disorder (GAD7)	Neuroticism
Raw data	2.0 ± 0.2219	0.034 ± 0.0027	0.172 ± 0.0066	0.160 ± 0.0137
BrainLM Pretrained	0.857 ± 0.1135	0.022 ± 0.0019	0.094 ± 0.0079	0.086 ± 0.0047
BrainLM fine-tuned	**0.485 ± 0.0252**	**0.018 ± 0.0008**	**0.074 ± 0.0053**	**0.072 ± 0.0049**

assessment of cognitive health and neural disorders. Ongoing work is exploring clinical prediction across a broader range of psychiatric, neurological, and neurodegenerative conditions.

4.3 Prediction of Future Brain States

To evaluate whether BrainLM can capture spatiotemporal dynamics, we assessed its performance in extrapolating to future brain states. A subset of UKB data was used to fine-tune the model to predict parcel activities at future timepoints.

During fine-tuning, BrainLM was given 180 timestep sequences and trained to forecast the subsequent 20 timesteps. We compared against baseline models including LSTMs, ODEnets, and a non-pretrained version of BrainLM.

As shown in Figure 4, the fine-tuned BrainLM model significantly outperformed other approaches in next timestep prediction on both UKB and HCP test sets. The non-pretrained BrainLM also struggled, confirming the benefits of pretraining for this task. Quantitatively, fine-tuned BrainLM achieved the lowest error across nearly all fore-

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.12.557460; this version posted September 13, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

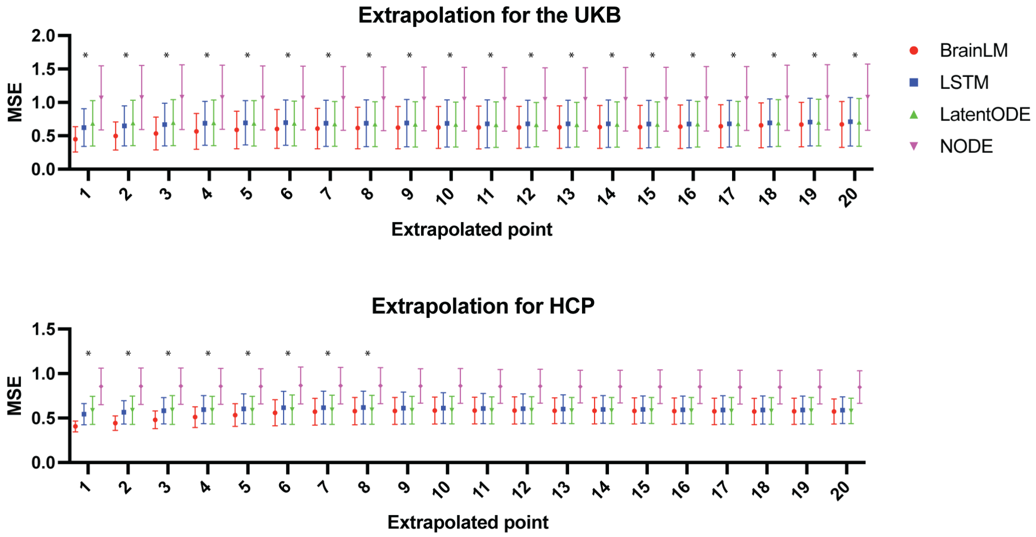

Figure 4: BrainLM outperforms other models in extrapolating future brain states. Models were trained to predict parcel activity 20 timesteps beyond observed context data. The plot shows the mean squared error per timestep on held-out UKB and HCP recordings. BrainLM demonstrates significantly lower error in forecasting near-future brain states. This highlights how pretraining enables BrainLM to effectively learn spatiotemporal fMRI dynamics. The time points for which BrainLM has significantly ($p < 0.05$) lower error than the other methods are identified with "*".

casted timesteps on UKB data, and significantly so for the first 8 timesteps on HCP recordings (see Table 3). This demonstrates BrainLM's strong ability to implicitly learn fMRI dynamics during pretraining.

Table 3: Quantitative evaluation of extrapolation performance. Models were tasked with forecasting parcel activity 40 timesteps beyond observed data from the UKB dataset. BrainLM shows the best performance across all metrics: higher (R^2) and Pearson correlation coefficients (R), and lower mean squared error (MSE) between predicted and true future states.

	UKB			HCP		
	R^2	R	MSE	R^2	R	MSE
BrainLM (fine-tuned)	**0.086**	**0.280**	**0.648**	**0.028**	**0.185**	**0.568**
BrainLM (w/o pre-training)	0.012	0.112	0.695	0.007	0.090	0.583
LSTM	-0.001	0.151	0.704	-0.020	0.049	0.598
Neural ODE	-0.577	0.001	1.083	-0.469	2.010e-4	0.857
Latent ODE	0.001	0.023	0.703	-0.003	-2.026e-4	0.588

4.4 Interpretability via Attention Analysis

A key advantage of BrainLM is interpretability. By visualizing the self-attention weights, we can extract insights about the model's representations. We averaged the attention that BrainLM's CLS token assigned to each parcel when encoding fMRI recordings. As shown in Figure 9, task recordings exhibited greater visual cortex attention compared to resting state, aligning with the visual stimuli presented during tasks. We also found differences in attention patterns between mild and severe depression (given by the PHQ9 scores), with more weighting on frontal and limbic areas for severe versus mild depression. These attention distributions agree with known functional alterations in depression [30, 31, 32]. Overall, the learned attention maps confirm that BrainLM captures clinically relevant variations in functional networks. By revealing which regions are prioritized during encoding, attention analysis provides useful interpretability for neuroscientific insights. For all other clinical variables, see supplementary Figure 9.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.12.557460; this version posted September 13, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

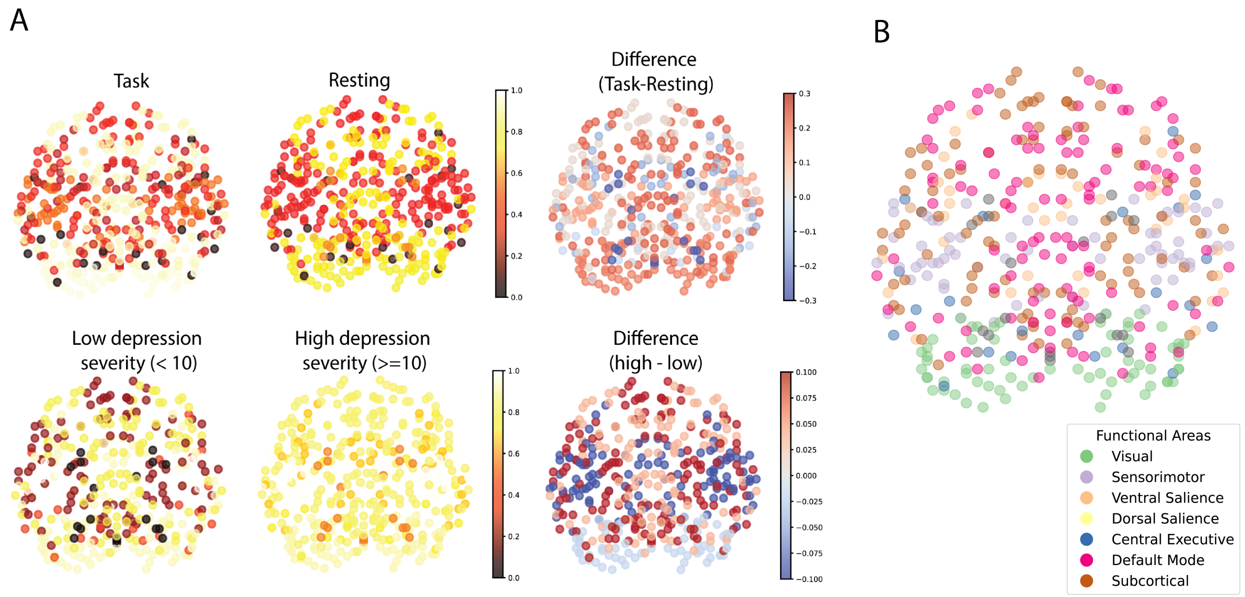

Figure 5: BrainLM attention maps reveal functional contrasts. A) Differences in parcel attention weights between task (left) and rest (right), and low and high depression scores (PHQ9). Task vs. resting state differences highlight changes in the visual cortex. Comparing the severity of depression, the difference highlights subcortical areas. B) Attention can localize parcels to 7 functional networks without supervision. This demonstrates BrainLM's ability to learn meaningful representations in an unsupervised manner.

Algorithm 1 *In Silico* Perturbation and Optimization with BrainLM

```
Require: Initial brain recording $X_{\text{Input}}$, initial perturbation parameters $\theta$, loss function (Mean Squared Error)
Ensure: Optimized perturbation parameters $\theta^*$
$\tilde{X} \leftarrow G(X_{\text{Input}}, \theta)$    ▷ Apply $G$ to perturb $X_{\text{Input}}$
$\hat{X} \leftarrow \text{BrainLM}(\tilde{X})$    ▷ Encode-decode with BrainLM
$\theta^* \leftarrow \arg\min_\theta \text{Loss}(X_{\text{Target}}, \text{BrainLM}(G(X_{\text{Input}}, \theta)))$    ▷ Optimize $\theta$ via gradient descent
return $\theta^*$
```

4.5 *In silico* Perturbation Analysis Reveals Functional Connectivity

We leveraged BrainLM's pre-trained representations to perform perturbation analysis and uncover functional brain networks. This was done by optimizing perturbations to make one fMRI recording mimic another, revealing key regional differences. Specifically, we optimized the perturbations to a resting-state recording that would alter its encoded representation (CLS token) to match that of a task-based recording. In Figure 7, we can observe the mean optimal perturbations for 1000 optimizations between resting state and task-based recordings. Here the perturbation focused on the visual cortex. This highlights the regions most impacted by the change from resting to task context, which involves visual processing. By analyzing recordings before and after optimization, we can elucidate which features are most influential in altering BrainLM's predicted state. This reveals the model's intrinsic knowledge of functional connections. Overall, this demonstrates how BrainLM can be prodded in silico to uncover meaningful functional relationships from its pre-trained representations, without additional tuning. We also performed the same procedure for recordings of different age groups (See Supplementary Figure 8).

4.6 Functional Network Prediction

We evaluated BrainLM's ability to segment parcels into intrinsic functional brain networks directly from fMRI activity patterns, without any network-based supervision. Parcels were categorized into 7 functional groups as defined in prior cortical parcellation work [cite]. The groups corresponded to visual, somatomotor, dorsal attention, ventral attention, limbic, frontoparietal, and default mode networks. On a held-out set of 1,000 UKB recordings, we compared different methods for classifying parcels into these 7 networks:

- k-NN classifier on raw parcel time series data

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.12.557460; this version posted September 13, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

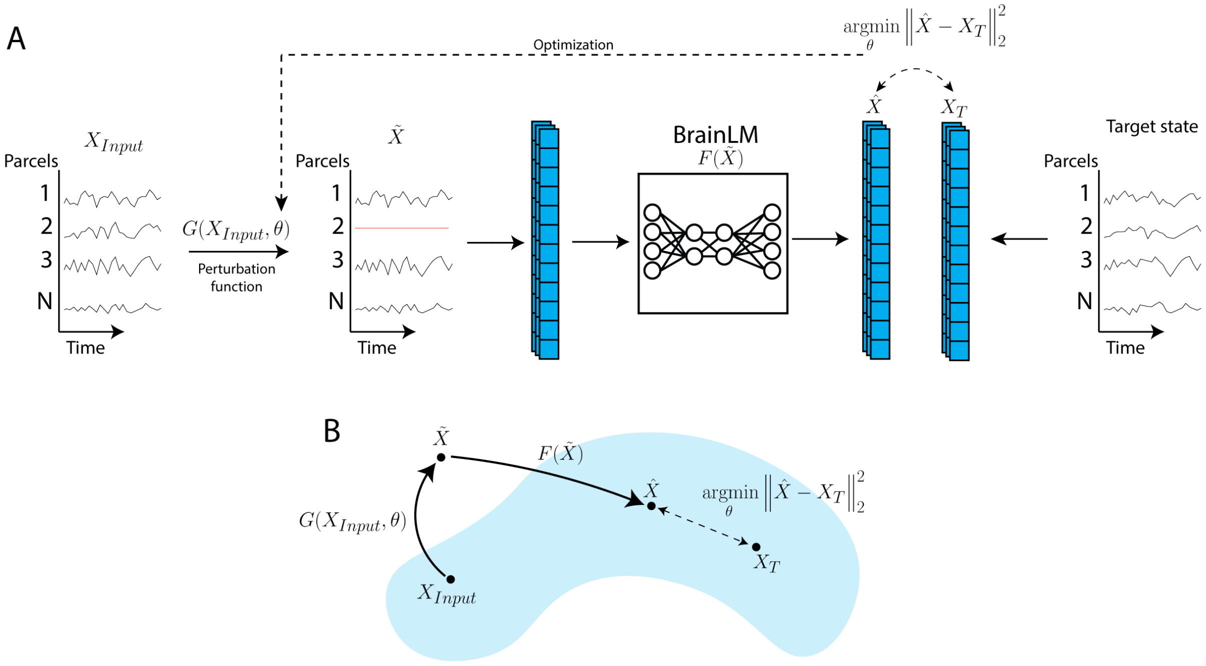

Figure 6: Overview of in silico perturbation approach. The goal of this procedure is to use the pre-trained BrainLM foundation model to simulate the effect of perturbations and find the modifications that result in a desired brain state. To achieve this, A) we learn a perturbation function G, which given an input state X_{Input} and parameters θ produces a perturbed state $\tilde{X}$. To find the optimal θ, we encode and decode the output of G and minimize the loss between the decoded perturbed state $\hat{X}$ with the target state X_T. B) Schematic of the *in silico* brain perturbation algorithm in the model manifold space.

Table 4: Comparing methods for functional region identification. Parcels from 1000 UKB recordings were categorized into 7 regions without supervision. A kNN classifier on BrainLM's self-attention maps achieved the highest accuracy, outperforming alternatives using raw data and other representation learning techniques.

	Accuracy (%)
BrainLM (attention weights)	**58.8**
Raw Data	39.2
Variational Autoencoder	49.4
Graph Convolutional Network	25.9

- k-NN on parcel embeddings from a VAE. The VAE contained 3 encoding and 3 decoding layers and was trained to reconstruct time series with the same masking ratio and loss as BrainLM.

- k-NN on parcel embeddings from a 4-layer Graph Convolutional Network (GCN). The GCN was trained with the same masking-based objective to learn parcel representations.

- k-NN on BrainLM's self-attention weights between each parcel token and all other parcel tokens.

The classifiers were trained on 80% of the parcels from each recording and evaluated on the remaining 20%. BrainLM's attention-based approach significantly outperformed the other methods, achieving 58.8% parcel classification accuracy (see Table 4). The k-NN on GCN performed worst at 25.9%. These results demonstrate BrainLM's unsupervised learning of functional brain topography from pretraining alone. The self-attention maps contain meaningful information about network identity without ever being exposed to explicit labels during training.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.12.557460; this version posted September 13, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

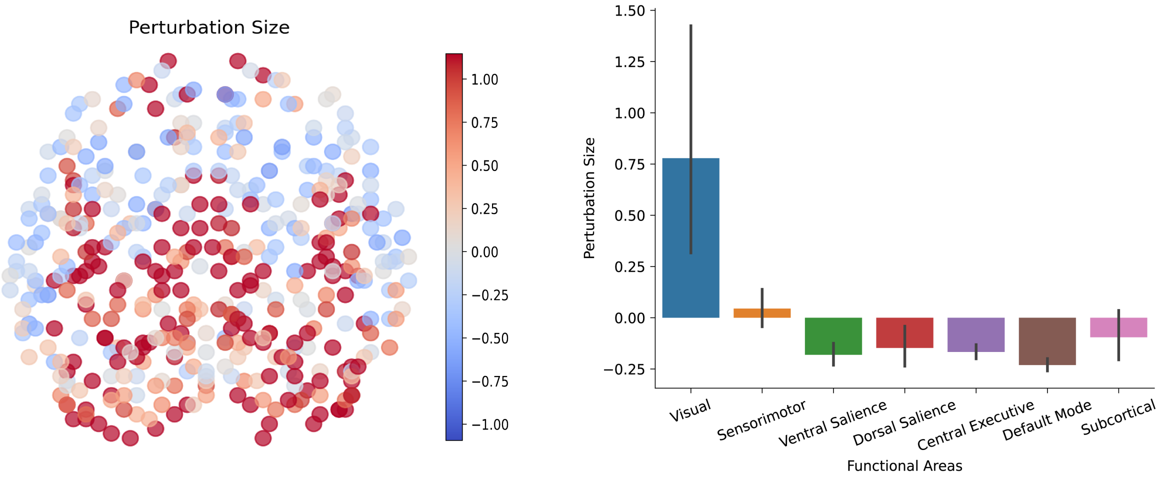

Figure 7: In silico perturbation of resting state to match task-based recordings reveals functional changes. The average magnitude of optimized perturbations to make resting state CLS tokens match target task CLS tokens. We find that the region with the largest predicted perturbation is the visual cortex, in line with expected functional changes between resting state and task-based recordings. This in silico perturbation approach demonstrates BrainLM's ability to simulate responses in a biologically meaningful manner.

5 Discussion

This work presents BrainLM, the first foundation model for functional MRI analysis. By leveraging self-supervised pretraining on 6,700 hours of brain activity recordings, BrainLM demonstrates versatile capabilities for modeling, predicting, and interpreting human brain dynamics.

A key innovation lies in BrainLM's generalizable representations of fMRI recordings. The model achieves high accuracy in reconstructing masked brain activity sequences, even generalizing to held-out distributions. This highlights the benefits of large-scale pretraining for learning robust encodings of spatiotemporal neural patterns. BrainLM also provides a powerful framework for biomarker discovery. By fine-tuning, brain dynamics can be decoded to predict clinical variables and psychiatric disorders better than baseline models. This could enable non-invasive assessment of cognitive health using resting-state fMRI alone.

Additionally, BrainLM offers new opportunities for computational modeling of brain function. We introduce a perturbation analysis technique that simulates fMRI responses by optimizing input changes that elicit a target model response. This demonstrates the utility of the approach for interpretability and causal discovery in a completely in silico manner without any intervention in live brain dynamics.

Finally, without any network-based supervision, BrainLM identifies intrinsic functional connectivity maps directly from pretraining, clustering parcels into known systems. This demonstrates how self-supervised objectives can extract organizational principles fundamental to the brain.

There remain exciting areas for future work. Multi-modal training could integrate fMRI with additional recording modalities, such as EEG and MEG, or different brain-wise information such as structural, functional, and genomic data. Probing a wider range of cognitive states and combined regularization may yield more generalizable representations. In future work, we could assess zero-shot classification on expanded functional atlases beyond the 7 networks used here.

Overall, BrainLM provides a springboard for accelerated research at the intersection of neuroscience and AI. We hope that this work spurs further development of foundation models that can help elucidate the intricate workings of the human brain.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.12.557460; this version posted September 13, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

6 Acknowledgement

This work is supported by the Wu Tsai Institute at Yale University. This research has been conducted using the UK Biobank Resource under Application Number 42826. Copyright © 2023, NHS England. Re-used with the permission of the NHS England [and/or UK Biobank]. All rights reserved. This research used data assets made available by National Safe Haven as part of the Data and Connectivity National Core Study, led by Health Data Research UK in partnership with the Office for National Statistics and funded by UK Research and Innovation (research which commenced between 1st October 2020 – 31st March 2021 grant ref $MC_PC_$20029; 1st April 2021 - 30th September 2022 grant ref $MC_PC_$20058).

References

[1] Bhedita J Seewoo, Alexander C Joos, and Kirk W Feindel. An analytical workflow for seed-based correlation and independent component analysis in interventional resting-state fmri studies. *Neuroscience research*, 165:26–37, 2021.

[2] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14453–14463, 2023.

[3] Milad Mozafari, Leila Reddy, and Rufin VanRullen. Reconstructing natural scenes from fmri patterns using bigbigan. In *2020 International joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2020.

[4] Furkan Ozcelik, Bhavin Choksi, Milad Mozafari, Leila Reddy, and Rufin VanRullen. Reconstruction of perceived images from fmri patterns and semantic brain exploration using instance-conditioned gans. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[6] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[7] Walter F Wiggins and Ali S Tejani. On the opportunities and risks of foundation models for natural language processing in radiology. *Radiology: Artificial Intelligence*, 4(4):e220119, 2022.

[8] Laurel J Orr, Karan Goel, and Christopher Ré. Data management opportunities for foundation models. In *CIDR*, 2022.

[9] Gengchen Mai, Weiming Huang, Jin Sun, Suhang Song, Deepak Mishra, Ninghao Liu, Song Gao, Tianming Liu, Gao Cong, Yingjie Hu, et al. On the opportunities and challenges of foundation models for geospatial artificial intelligence. *arXiv preprint arXiv:2304.06798*, 2023.

[10] Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature communications*, 8(1):15037, 2017.

[11] Sebastian Hoefle, Annerose Engel, Rodrigo Basilio, Vinoo Alluri, Petri Toiviainen, Maurício Cagy, and Jorge Moll. Identifying musical pieces from fmri data using encoding and decoding models. *Scientific reports*, 8(1):2266, 2018.

[12] Roman Beliy, Guy Gaziv, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. From voxels to pixels and back: Self-supervision in natural-image reconstruction from fmri. *Advances in Neural Information Processing Systems*, 32, 2019.

[13] Zijiao Chen, Jiaxin Qing, Tiange Xiang, Wan Lin Yue, and Juan Helen Zhou. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22710–22720, 2023.

[14] Zijiao Chen, Jiaxin Qing, and Juan Helen Zhou. Cinematic mindscapes: High-quality video reconstruction from brain activity. *arXiv preprint arXiv:2305.11675*, 2023.

[15] Emanuele Zappala, Antonio Henrique de Oliveira Fonseca, Josue Ortega Caro, and David van Dijk. Neural integral equations. *arXiv preprint arXiv:2209.15190*, 2022.

[16] Antonio H de O Fonseca, Emanuele Zappala, Josue Ortega Caro, and David van Dijk. Continuous spatiotemporal transformers. *arXiv preprint arXiv:2301.13338*, 2023.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.12.557460; this version posted September 13, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[17] Joel Ye and Chethan Pandarinath. Representation learning for neural population activity with neural data transformers. *arXiv preprint arXiv:2108.01210*, 2021.

[18] Edgar Y Walker, Fabian H Sinz, Erick Cobos, Taliah Muhammad, Emmanouil Froudarakis, Paul G Fahey, Alexander S Ecker, Jacob Reimer, Xaq Pitkow, and Andreas S Tolias. Inception loops discover what excites neurons most using deep predictive models. *Nature neuroscience*, 22(12):2060–2065, 2019.

[19] Josue Ortega Caro, Yilong Ju, Ryan Pyle, Sourav Dey, Wieland Brendel, Fabio Anselmi, and Ankit Patel. Local convolutions cause an implicit bias towards high frequency adversarial examples. *arXiv preprint arXiv:2006.11440*, 2020.

[20] Zhe Li, Josue Ortega Caro, Evgenia Rusak, Wieland Brendel, Matthias Bethge, Fabio Anselmi, Ankit B Patel, Andreas S Tolias, and Xaq Pitkow. Robust deep learning object recognition models rely on low frequency information in natural images. *PLOS Computational Biology*, 19(3):e1010932, 2023.

[21] Daniel M Bear, Kevin Feigelis, Honglin Chen, Wanhee Lee, Rahul Venkatesh, Klemen Kotar, Alex Durango, and Daniel LK Yamins. Unifying (machine) vision via counterfactual world modeling. *arXiv preprint arXiv:2306.01828*, 2023.

[22] Karla L Miller, Fidel Alfaro-Almagro, Neal K Bangerter, David L Thomas, Essa Yacoub, Junqian Xu, Andreas J Bartsch, Saad Jbabdi, Stamatios N Sotiropoulos, Jesper LR Andersson, et al. Multimodal population brain imaging in the uk biobank prospective epidemiological study. *Nature neuroscience*, 19(11):1523–1536, 2016.

[23] Jennifer Stine Elam, Matthew F Glasser, Michael P Harms, Stamatios N Sotiropoulos, Jesper LR Andersson, Gregory C Burgess, Sandra W Curtiss, Robert Oostenveld, Linda J Larson-Prior, Jan-Mathijs Schoffelen, et al. The human connectome project: a retrospective. *NeuroImage*, 244:118543, 2021.

[24] Gholamreza Salimi-Khorshidi, Gwenaëlle Douaud, Christian F Beckmann, Matthew F Glasser, Ludovica Griffanti, and Stephen M Smith. Automatic denoising of functional mri data: combining independent component analysis and hierarchical fusion of classifiers. *Neuroimage*, 90:449–468, 2014.

[25] Chadi G Abdallah. Brain networks associated with covid-19 risk: Data from 3662 participants. *Chronic Stress*, 5:24705470211066770, 2021.

[26] Samaneh Nemati, Teddy J Akiki, Jeremy Roscoe, Yumeng Ju, Christopher L Averill, Samar Fouda, Arpan Dutta, Shane McKie, John H Krystal, JF William Deakin, et al. A unique brain connectome fingerprint predates and predicts response to antidepressants. *IScience*, 23(1), 2020.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[28] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[29] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[30] Diego A Pizzagalli and Angela C Roberts. Prefrontal cortex and depression. *Neuropsychopharmacology*, 47(1):225–246, 2022.

[31] NL Johnston-Wilson, CD Sims, JP Hofmann, L Anderson, AD Shore, EF Torrey, and Robert H Yolken. Disease-specific alterations in frontal cortex brain proteins in schizophrenia, bipolar disorder, and major depressive disorder. *Molecular psychiatry*, 5(2):142–149, 2000.

[32] Te-Jen Lai, Martha E Payne, Christopher E Byrum, David C Steffens, and K Ranga R Krishnan. Reduction of orbital frontal cortex volume in geriatric depression. *Biological psychiatry*, 48(10):971–975, 2000.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.12.557460; this version posted September 13, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A Supplementary Figures

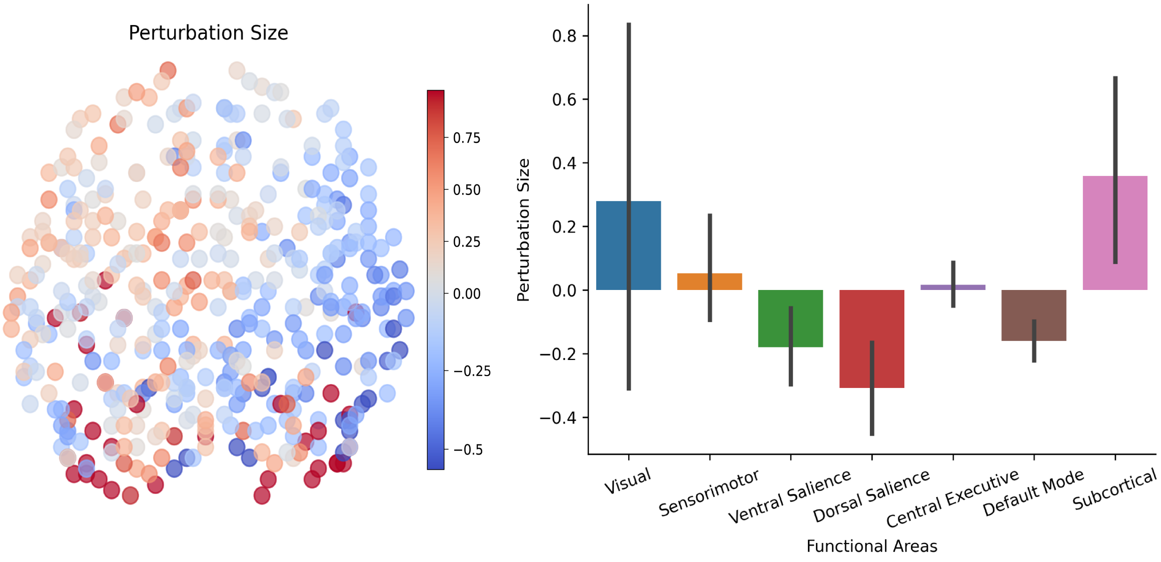

Figure 8: Average perturbation to match the CLS token of a below 65-year-old patient recording to a target CLS token from above 65-year-old patient recording. The average was computed over 1000 patients of the validation set. We found large changes in the visual cortex and somatosensory cortex. We also observe these differences are asymmetric across hemispheres between the two CLS tokens.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.12.557460; this version posted September 13, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

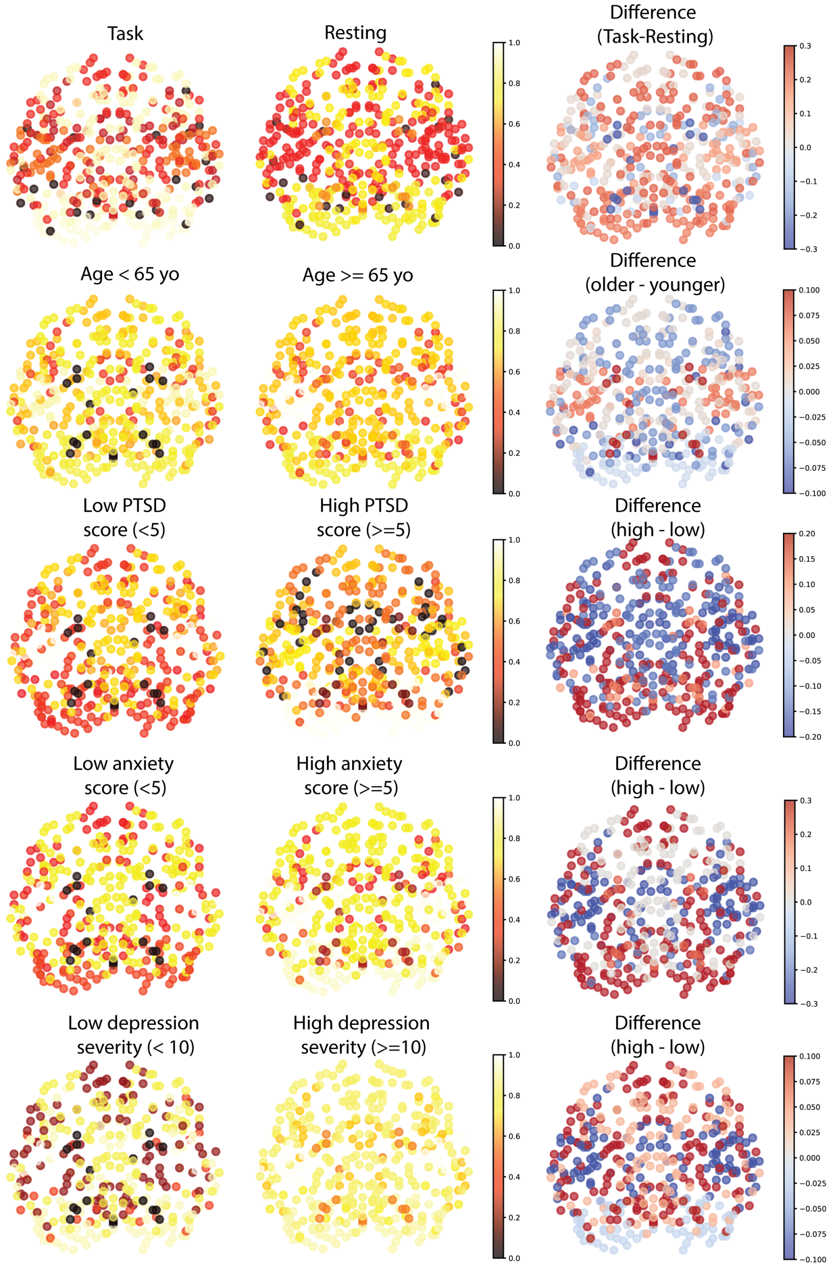

Figure 9: Attention visualization across 424 parcels. Each parcel is localized by its X and Y position and it is colored by the attention intensity with respect to the CLS token.