Sources and Sinks in Functional Brain Networks

Keith Dillon

November 24, 2025

Abstract

In this paper we measured the degree to which functional brain activity signals in resting-state BOLD fMRI could be predicted using a scalable self-supervised learning approach. Such predictability is at the core of methods ranging from using simple correlations to identify edges of a network to large foundation models of brain activity. To provide interpretability we trained a linear autoencoder to predict the fMRI time series using its past history. We first analyzed the variation in prediction error between brain regions, finding it highest in visual and motor regions and lowest in limbic and subcortical regions. We then investigated regions that were involved in predictive activity but were less able to be predicted themselves, which we call sources, or else were able to be predicted but contributed less to predictability of other regions, which we call sinks. We found higher-order sensory regions to be the most prominent sources, while motor coordination and primary visual regions were the most prominent sinks.

Introduction

A wide range of methods have been applied to functional imaging data to estimate information of scientific or clinical interest. Blood oxygenation level dependent functional magnetic resonance imaging (BOLD fMRI) [12] has been especially popular in such research as it is noninvasive and provides relatively high spatial resolution. Initial research identified regions where brain activity correlated with different tasks or self-reported brain states. But it became apparent that many complex brain states, as well as mental illnesses such as schizophrenia [14], could not be pinpointed as a deficit in a specific brain region. The subsequent assumption has been that these might be described as resulting from coordinated activity, referred to as brain networks. Large datasets have been generated by several collaborations, such as the Human Connectome Project [13], which sought to identify functional brain networks using data-driven methods. Many fundamental methods for performing such an estimate can be formulated based on a regression problem of describing the activity in one region using the activity in other regions, which can be viewed as a form of self-supervised learning. Examples include such broad categories as spectral network methods [4], Gaussian graphical models [2], and partial correlation [3].

More recently, deep learning methods from natural language processing (NLP) have been adopted. Such methods are commonly trained by self-supervised learning methods where the model is optimized to predict masked portions of the data given other unmasked portions, over a vast dataset. Such models essentially learn a joint probability distribution of the data. In NLP they are called foundation models, as they can be subsequently used for other tasks by making conditional inferences on this distribution. When applied to fMRI data, this self-supervised approach is again a regression problem. An increasing number of recent works use such approaches [15].

Whether the goal is to extract a network description to be analyzed and interpreted directly, or create a predictive model that can be used in subsequent tasks, such regression-based models all restrict their information content to that which can be predicted from activity. In natural language this restriction seems more valid as the goal of language is to communicate information. There is no such driving force behind functional imaging; its information content is simply that which current technology manages to measure. Latent information is viewed statistically, as in some random chance to arise depending on previous outputs. The purpose of this paper is to analyze this information in more structural detail.

Results

We used the 100 unrelated subjects dataset from the Human Connectome Project [13]. Each scan contains functional Magnetic Resonance Imaging (fMRI) data with 96,854 time series, each of which is sampled from a single voxel in the brain with 1,200 time samples each. The data was preprocessed as described in [6] and was subsequently filtered spatially using a 5 mm kernel with the Connectome Workbench [7]. Finally, each time series was standardized to zero mean and unit variance. Cortical regions were parcellated using the Human Connectome Project Multimodal Parcellation (MMP) Atlas [5], which resulted in 379 regions of interest (ROI) when combined with subcortical regions. We implemented a linear autoencoder model that predicts the total fMRI image at a point in time using the images at prior times. We tested a range of time window sizes and hidden-layer sizes and found a window of 12 seconds (6 samples) and a 128-node hidden layer provided the best prediction. As this resulted in roughly 291 thousand parameters, we used PyTorch [8] to implement it as a neural network model with two linear layers and trained it using stochastic gradient descent with a least-squares loss function.

An example of the result is shown in Fig. 1, giving the ROI for which the predictions are best and worst in terms of mean-squared error (MSE). The best predicted region is the cortical ROI labeled "Visual2-04_R" in

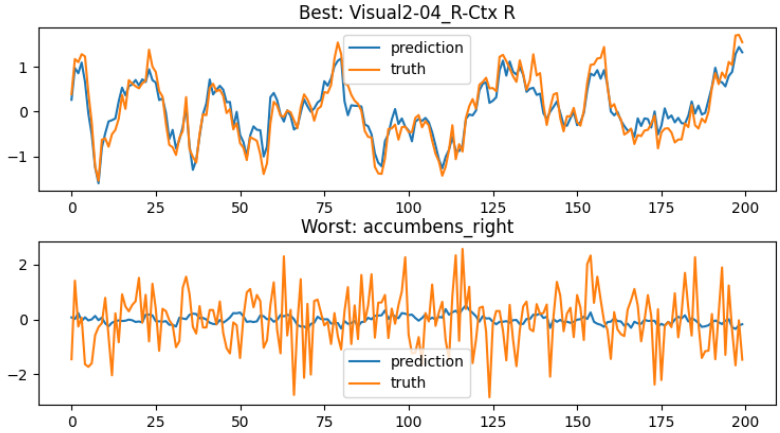

Figure 1: Time series for ROI with least and most MSE.

the visual cortex. The worst predicted region is the right accumbens in the basal ganglia. See Table 1 in the appendix for the 30 best and worst MSE regions.

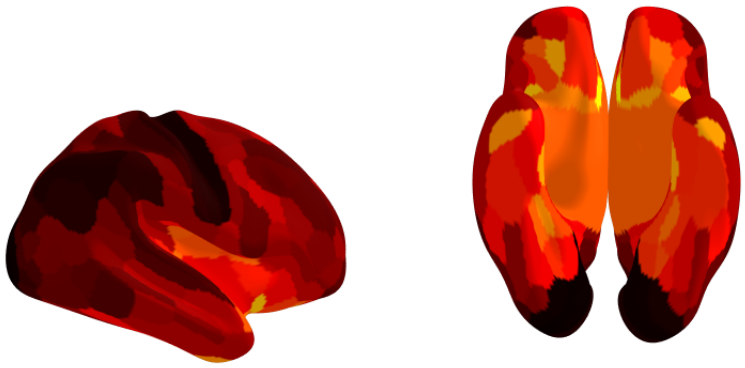

Figure 2: Surface map of MSE, right and bottom views; hotter (lighter) colors are higher MSE.

Fig. 2 shows a surface map of the MSE for different regions for right and bottom views of the brain. Sensory and somatomotor regions clearly dominate the low-MSE regions, while high-MSE regions are involved in reward processing, emotional regulation, decision-making, motor control, and sensory integration, among other functions. Focusing on the cortical regions with high MSE, we find regions involved in cognitive control,

emotional processing, language function, executive decision-making, self-referential thought, and sensory-motor integration.

In Fig. 1 we also note that the worst-predicted region has a much higher frequency signal, which may reasonably be expected to be more difficult to predict versus a region with slower activity. As a way of normalizing against such effects, we next considered the relative predictability of regions by comparing how much they could be predicted versus the degree to which they were useful in predicting other regions. We defined the predictability of an ROI as the absolute sum of incoming weights to the node predicting the ROI (the in-degree). We similarly defined the degree to which an ROI contributed to prediction of others as the absolute sum of weights using the ROI to predict others (out-degree).

Of particular interest were those ROI that were useful to predict other ROI but were poorly predicted themselves using others, which we called sources. These might be viewed as disproportionately receiving inputs from some latent source, such as incoming connections to the brain via sensory information. We estimated these by taking the ratio of absolute out-degree over absolute in-degree. Conversely, we defined sinks as the inverse of this ratio, meaning ROI that were well predicted by signals from elsewhere but contributed little themselves to further signal prediction. These might be viewed as connecting outputs from the brain to the body, among other possible functions. The ROI with the strongest ratios are visualized in Fig. 3. Sources and sinks are listed in more detail in the appendix.

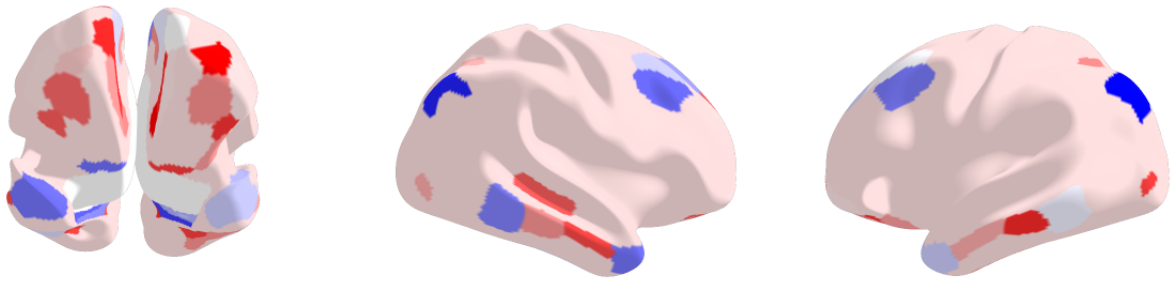

Figure 3: Visualization of strongest sources and sinks; red is higher row/col (sources), blue is higher col/row (sinks).

Sources include regions involved in visual and auditory processing, executive function, language, motor control, attention, and sensory integration. Sinks include regions that are involved in motor coordination, visual processing, executive function, default mode network activity, and cognitive control. Note that some functions are included in both lists, and well-known networks such as the default mode network [9] include both sources and sinks.

Discussion

We provided a basic method to quantify what information is missing from a model trained in a self-supervised manner and provided an analysis of its spatial variation using real fMRI data. The results appear reasonable, as sensory and motor regions involved in information incoming to the brain appeared as sources. Similarly, sinks prominently included regions involved in providing external outputs from the brain, such as the cerebellum. More interestingly, higher cognitive functions involved both sources and sinks, which may have multiple explanations. One is that the mental states involving these regions are indeed terminal also, either appearing spontaneously to lead to subsequent states, or else resulting from some past states but perhaps extinguishing that process in some competitive decision process.

It is also possible that higher cognitive functions involve complex nonlinear behavior that is harder to predict, even when the available information might be there. We attempted to employ more complex models, such as using nonlinear activation functions and more layers in a deep neural network. However, we were not able to achieve improvement gains over the linear model, perhaps due to the high noise level in the data, especially relative to the dataset size.

As increasingly large datasets become available, other researchers have been utilizing more complex modern

deep learning models. Examples include [11] and [1], which use self-supervised learning methods from modern natural language processing (NLP). Explainable artificial intelligence (XAI) methods [10] might be adapted to provide structural interpretations of sources and sinks in such models. Though it should be noted there is no reason to expect current NLP architectures would prove particularly effective for the very noisy and still relatively much smaller fMRI datasets compared to NLP. Also, current published methods often use large sets of task-fMRI data while not utilizing the task information itself, which would help reduce the missing latent information. And these publications lack a linear baseline for comparison, instead focusing on usefulness of the model as forming representations useful in so-called downstream tasks.

References

[1] Josue Ortega Caro, Antonio Henrique de Oliveira Fonseca, Syed A. Rizvi, Matteo Rosati, Christopher Averill, James L. Cross, Prateek Mittal, Emanuele Zappala, Rahul Madhav Dhodapkar, Chadi Abdallah, and David van Dijk. BrainLM: A foundation model for brain activity recordings. October 2023.

[2] Keith Dillon. Clustering Gaussian Graphical Models. *arXiv:1910.02342 [cs, stat]*, October 2019. arXiv: 1910.02342.

[3] Keith Dillon. Efficient Partitioning of Partial Correlation Networks. In Andrea Torsello, Luca Rossi, Marcello Pelillo, Battista Biggio, and Antonio Robles-Kelly, editors, *Structural, Syntactic, and Statistical Pattern Recognition*, Lecture Notes in Computer Science, pages 174–183, Cham, 2021. Springer International Publishing.

[4] Keith Dillon and Yu-Ping Wang. Resolution-based spectral clustering for brain parcellation using functional MRI. *Journal of Neuroscience Methods*, 335:108628, April 2020.

[5] Matthew F. Glasser, Timothy S. Coalson, Emma C. Robinson, Carl D. Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F. Beckmann, Mark Jenkinson, Stephen M. Smith, and David C. Van Essen. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, August 2016. Publisher: Nature Publishing Group.

[6] Matthew F. Glasser, Stephen M. Smith, Daniel S. Marcus, Jesper L. R. Andersson, Edward J. Auerbach, Timothy E. J. Behrens, Timothy S. Coalson, Michael P. Harms, Mark Jenkinson, Steen Moeller, Emma C. Robinson, Stamatios N. Sotiropoulos, Junqian Xu, Essa Yacoub, Kamil Ugurbil, and David C. Van Essen. The Human Connectome Project's neuroimaging approach. *Nature Neuroscience*, 19(9):1175–1187, September 2016. Publisher: Nature Publishing Group.

[7] Daniel S. Marcus, Michael P. Harms, Abraham Z. Snyder, Mark Jenkinson, J. Anthony Wilson, Matthew F. Glasser, Deanna M. Barch, Kevin A. Archie, Gregory C. Burgess, Mohana Ramaratnam, Michael Hodge, William Horton, Rick Herrick, Timothy Olsen, Michael McKay, Matthew House, Michael Hileman, Erin Reid, John Harwell, Timothy Coalson, Jon Schindler, Jennifer S. Elam, Sandra W. Curtiss, and David C. Van Essen. Human Connectome Project informatics: Quality control, database services, and data visualization. *NeuroImage*, 80:202–219, October 2013.

[8] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[9] Marcus E. Raichle. The brain's default mode network. *Annual Review of Neuroscience*, 38:433–447, July 2015.

[10] Nys Tjade Siegel, James H. Cole, Mohamad Habes, Stefan Haufe, Kerstin Ritter, and Marc-André Schulz. Explainable AI Methods for Neuroimaging: Systematic Failures of Common Tools, the Need for Domain-Specific Validation, and a Proposal for Safe Application, August 2025. arXiv:2508.02560 [cs].

[11] Armin W. Thomas, Christopher Ré, and Russell A. Poldrack. Self-Supervised Learning of Brain Dynamics from Broad Neuroimaging Data. October 2022.

[12] Martijn P. van den Heuvel and Hilleke E. Hulshoff Pol. Exploring the brain network: A review on resting-state fMRI functional connectivity. *European Neuropsychopharmacology*, 20(8):519–534, August 2010.

[13] D. C. Van Essen, K. Ugurbil, E. Auerbach, D. Barch, T. E. J. Behrens, R. Bucholz, A. Chang, L. Chen, M. Corbetta, S. W. Curtiss, S. Della Penna, D. Feinberg, M. F. Glasser, N. Harel, A. C. Heath, L. Larson-Prior, D. Marcus, G. Michalareas, S. Moeller, R. Oostenveld, S. E. Petersen, F. Prior, B. L. Schlaggar, S. M. Smith, A. Z. Snyder, J. Xu, E. Yacoub, and WU-Minn HCP Consortium. The Human Connectome Project: a data acquisition perspective. *NeuroImage*, 62(4):2222–2231, October 2012.

[14] Heather C. Whalley, Enrico Simonotto, William Moorhead, Andrew McIntosh, Ian Marshall, Klaus P. Ebmeier, David G. C. Owens, Nigel H. Goddard, Eve C. Johnstone, and Stephen M. Lawrie. Functional Imaging as a Predictor of Schizophrenia. *Biological Psychiatry*, 60(5):454–462, September 2006.

[15] Xinliang Zhou, Chenyu Liu, Zhisheng Chen, Kun Wang, Yi Ding, Ziyu Jia, and Qingsong Wen. Brain Foundation Models: A Survey on Advancements in Neural Signal Processing and Brain Discovery, March 2025. ADS Bibcode: 2025arXiv250300580Z.

Appendix

This appendix gives tables listing the ROI with highest and lowest MSE, as well as with highest source or sink metrics.

Worst MSE Regions	**Best MSE Regions**
accumbens_right	Visual2 04 R Ctx R
accumbens_left	Visual2 03 R Ctx R
pallidum_right	Visual2 05 R Ctx R
pallidum_left	Visual2 30 L Ctx L
amygdala_left	Default 70 L Ctx L
amygdala_right	Visual2 31 L Ctx L
Default35 R Ctx R	Somatomotor 01 R Ctx R
Default74 L Ctx L	Somatomotor 21 L Ctx L
diencephalon_left	Visual2 32 L Ctx L
diencephalon_right	Visual1 04 L Ctx L
OrbitoAffective02 R Ctx R	Visual1 01 R Ctx R
OrbitoAffective05 L Ctx L	Frontoparietal 48 L Ctx L
CinguloOpercular37 L Ctx L	Default 32 R Ctx R
thalamus_right	Frontoparietal 24 R Ctx R
Frontoparietal16 R Ctx R	Somatomotor 31 L Ctx L
Default22 R Ctx R	Somatomotor 11 R Ctx R
Default61 L Ctx L	Somatomotor 22 L Ctx L
Language23 L Ctx L	Visual2 07 R Ctx R
thalamus_left	Somatomotor 02 R Ctx R
OrbitoAffective06 L Ctx L	Default 31 R Ctx R
CinguloOpercular36 L Ctx L	Default 69 L Ctx L
Somatomotor39 L Ctx L	Visual1 03 R Ctx R
Language09 R Ctx R	CinguloOpercular 51 L Ctx L
Default33 R Ctx R	CinguloOpercular 24 R Ctx R
Default75 L Ctx L	Visual2 34 L Ctx L
OrbitoAffective03 R Ctx R	Somatomotor 09 R Ctx R
Frontoparietal04 R Ctx R	Frontoparietal 23 R Ctx R
Default36 R Ctx R	Somatomotor 29 L Ctx L
Frontoparietal41 L Ctx L	Dorsal Attention 10 R Ctx R
OrbitoAffective01 R Ctx R	Visual1 06 L Ctx L

Table 1: Worst and Best MSE Regions

ROI	MSE	Row Sum	Col Sum	Row/Col
Visual2-53 L Ctx L	22580	11.29	7.56	1.49
Visual2-26 R Ctx R	21003	10.47	7.18	1.46
Visual2-21 R Ctx R	17749	10.42	7.52	1.39
Auditory-14 L Ctx L	16811	9.67	6.99	1.38
Frontoparietal-16 R Ctx R	28870	10.57	7.66	1.38
Visual2-48 L Ctx L	18938	10.59	7.79	1.36
Visual2-20 R Ctx R	18735	10.02	7.51	1.33
Frontoparietal-41 L Ctx L	25606	9.44	7.10	1.33
Language-09 R Ctx R	27073	8.95	6.73	1.33
Somatomotor-16 R Ctx R	18731	10.05	7.59	1.32
Default-44 L Ctx L	18689	10.30	7.80	1.32
Auditory-11 L Ctx L	25446	8.94	6.82	1.31
Frontoparietal-39 L Ctx L	13306	11.62	8.99	1.29
Default-75 L Ctx L	26398	10.07	7.79	1.29
Dorsal-Attention-20 L Ctx L	19066	9.40	7.30	1.29
Cingulo-Opercular-56 L Ctx L	23381	9.56	7.44	1.29
Default-34 R Ctx R	12334	10.45	8.16	1.28
Auditory-05 R Ctx R	20075	8.89	6.94	1.28
thalamus_left	28292	7.58	5.93	1.28
Visual2-15 R Ctx R	10456	11.25	8.83	1.27
Somatomotor-17 R Ctx R	23978	9.05	7.12	1.27
Visual2-28 L Ctx L	13161	10.04	7.93	1.27
Somatomotor-24 L Ctx L	15216	9.78	7.73	1.26
amygdala_left	39952	7.91	6.26	1.26
Frontoparietal-25 R Ctx R	15630	11.05	8.77	1.26
Default-03 R Ctx R	21360	9.30	7.43	1.25
Auditory-08 L Ctx L	19227	9.50	7.60	1.25
Language-20 L Ctx L	13894	10.34	8.28	1.25
Default-05 R Ctx R	19074	9.93	7.96	1.25
Default-46 L Ctx L	14162	11.09	8.90	1.25

Table 2: Top source ROI and region metrics

ROI	MSE	Row Sum	Col Sum	Col/Row
cerebellum_left	7964	9.61	18.50	1.93
cerebellum_right	6868	9.26	17.56	1.90
Visual1-04 L Ctx L	4501	9.51	16.68	1.75
Visual1-01 R Ctx R	4510	9.48	16.10	1.70
Frontoparietal-30 L Ctx L	6950	9.94	16.29	1.64
Somatomotor-21 L Ctx L	4244	8.02	12.87	1.60
Frontoparietal-02 R Ctx R	7275	10.19	16.34	1.60
Default-70 L Ctx L	3683	9.98	15.94	1.60
Frontoparietal-48 L Ctx L	4539	9.65	15.03	1.56
Frontoparietal-24 R Ctx R	4760	9.82	15.04	1.53
Somatomotor-01 R Ctx R	4235	8.71	13.27	1.52
Frontoparietal-29 L Ctx L	10951	6.85	10.22	1.49
Default-32 R Ctx R	4652	10.22	15.24	1.49
Frontoparietal-01 R Ctx R	10642	7.51	11.12	1.48
Visual2-31 L Ctx L	3700	8.00	11.79	1.47
Visual2-04 R Ctx R	3128	8.09	11.60	1.43
Default-62 L Ctx L	16709	6.61	9.31	1.41
Visual2-30 L Ctx L	3592	8.17	11.41	1.40
Cingulo-Opercular-12 R Ctx R	9220	8.66	12.02	1.39
Default-69 L Ctx L	5349	9.67	13.24	1.37
Dorsal-Attention-22 L Ctx L	6113	10.49	14.27	1.36
Cingulo-Opercular-51 L Ctx L	5578	9.93	13.28	1.34
Default-31 R Ctx R	5223	9.97	13.25	1.33
Default-24 R Ctx R	18514	6.83	9.06	1.33
Dorsal-Attention-04 R Ctx R	10751	8.62	11.40	1.32
Ventral-Multimodal-03 L Ctx L	20461	5.94	7.83	1.32
Frontoparietal-07 R Ctx R	9089	9.29	12.11	1.30
Visual1-06 L Ctx L	6104	9.19	11.93	1.30
Orbito-Affective-01 R Ctx R	25551	5.85	7.57	1.29
Orbito-Affective-03 R Ctx R	26229	5.91	7.63	1.29

Table 3: Top sink ROI and region metrics