MAX PLANCK INSTITUTE FOR SOFTWARE SYSTEMS

PRINCETON UNIVERSITY

Machine Learning Department

Computational Biology Department

Neuroscience Institute

Carnegie Mellon University

Same Cause; Different Effects in the Brain

Mariya Toneva* — Princeton, MPI-SWS
Jennifer Williams* — CMU
Anand Bollu — CMU
Christoph Dann — Google
Leila Wehbe — CMU

Summary

Background: To map information processing in the brain, researchers use encoding models to evaluate if stimulus properties predict brain data.

Gap in the field: Naturalistic stimuli make it difficult to infer what stimulus properties affect each brain zone because the stimuli are multivariate and often high-dimensional.

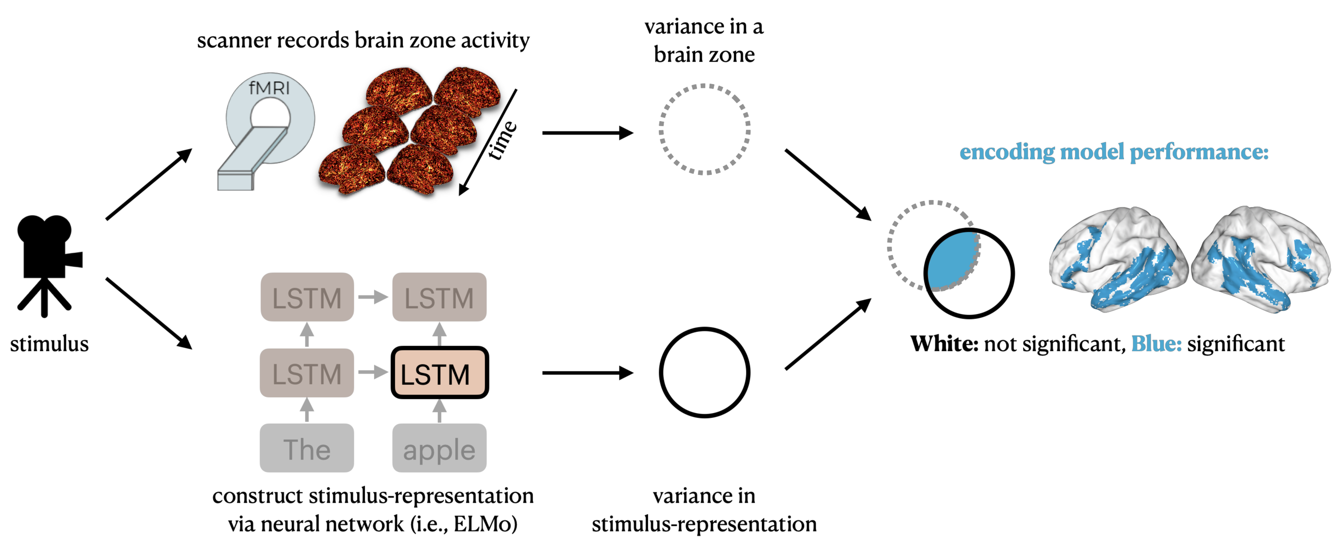
Main contribution: Enable researchers to infer if a stimulus affects two brain zones in the same way by proposing an inference framework that includes two new metrics.

Validation:

Simulations show that the proposed metrics provide new insights beyond current brain mapping techniques.

Consistent inferences across 2 naturalistic fMRI datasets, acquired from different subjects, labs, and stimuli.

Motivation



Encoding model:

$$Y_i = g_i(X) + \epsilon_i$$

$Y_i \in \mathbb{R}$ observation in zone i

$X \in \mathbb{R}^d$ stimulus-representation

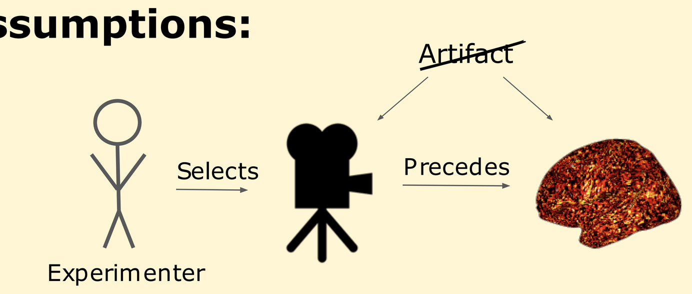
$g_i(X) = \langle X, \theta_i \rangle$ stimulus effect

Claim: encoding model cannot infer if stim. properties affect 2 brain zones in the same way

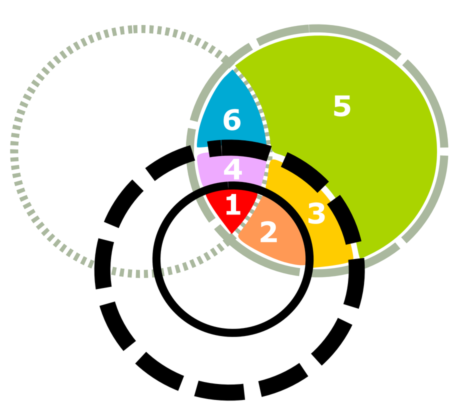
Causal interpretation:

reveals which brain zones affected by stimulus properties captured by stimulus-representation [1]

Assumptions:



1+4 similar effect
2+3 different effect
1 similar effect of stim. properties captured by ELMo
4 similar effect of stim. properties missing from ELMo
2 different effect of stim. properties captured by ELMo
3 different effect of stim. properties missing from ELMo
5 different noise
6 similar noise



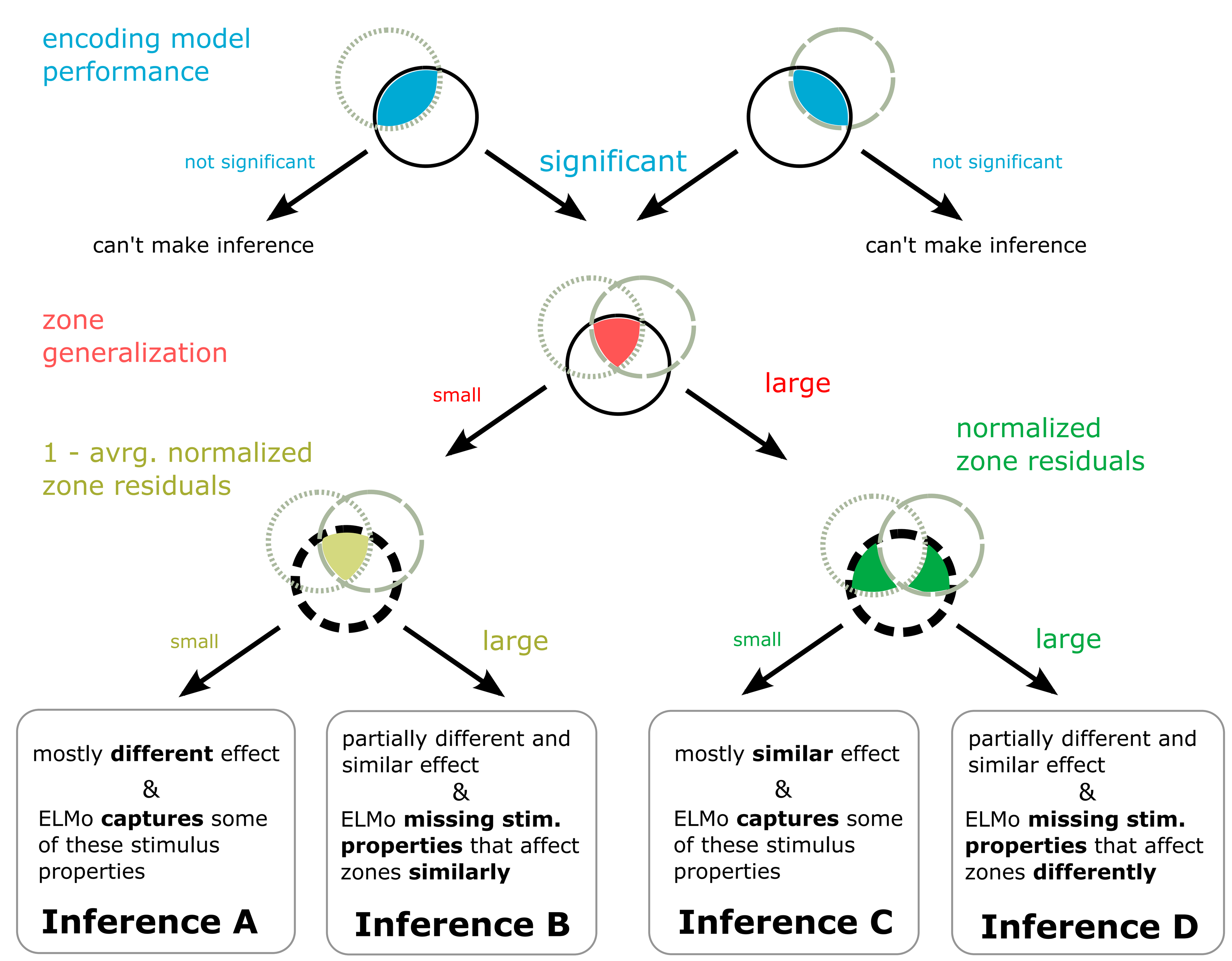
variance in a brain zone; variance in a second brain zone; variance in stimulus representation (e.g. ELMo); variance in stimulus

Code: github.com/brainML/stim-effect

Inference Framework

Zone generalization: how similarly two zones are affected by stim. properties in the stimulus-representation (area 1)

Zone residuals: capture any stimulus effect that is not shared between two zones (areas 2 & 3)



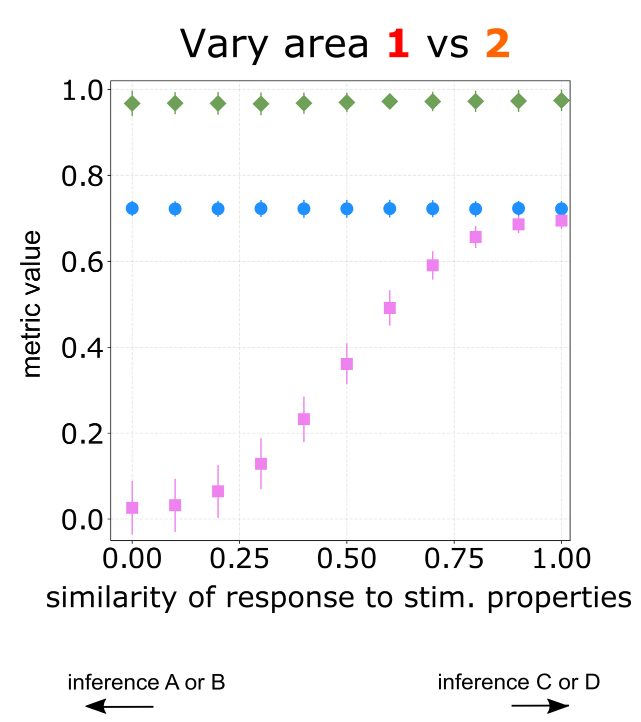
Metrics Implementation & Validation

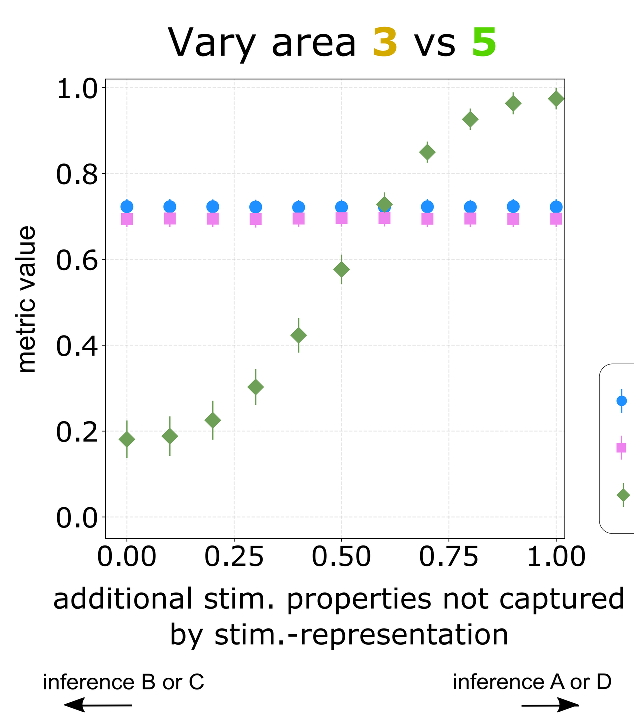
encoding model performance $(\text{zone}_i) = \text{corr}(\widehat{Y}_i, Y_i)$ commonly used, e.g. [2-3]

zone generalization $(\text{zone}_i, \text{zone}_j) = \text{corr}(\widehat{Y}_i, Y_j)$ inspired by [4-5]

zone residuals $(\text{zone}_i, \text{zone}_j) = \frac{1}{M^2 - M} \sum_{S,T,S \neq T} \text{corr}(R_{i-j,S}, R_{i-j,T}),$ inspired by [6]

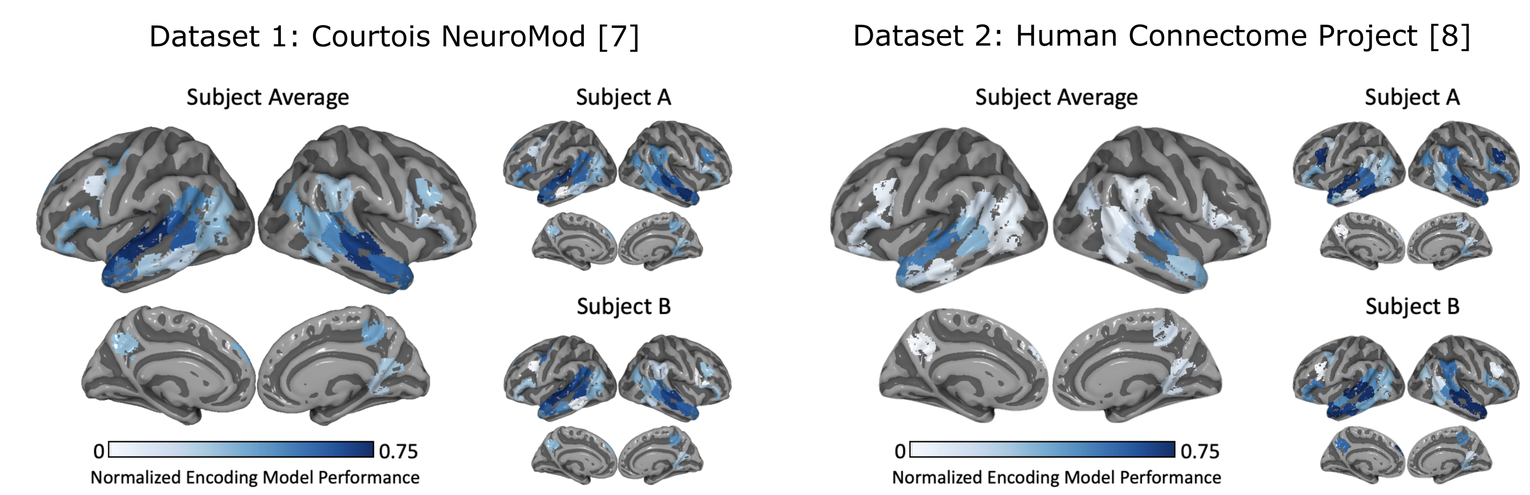
$$R_{i-j,P} = Y_{i,P} - Y_{j,P}\beta_P^{ij}$$



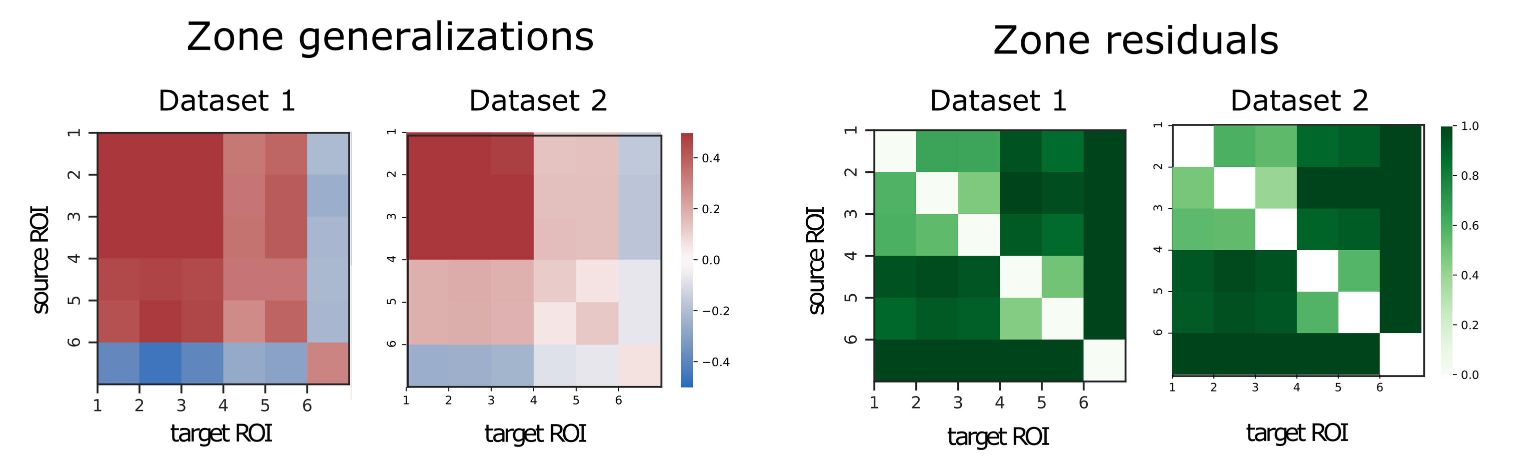


Both types of new metrics needed to make one of the **4 inferences**

Results on 2 fMRI Datasets

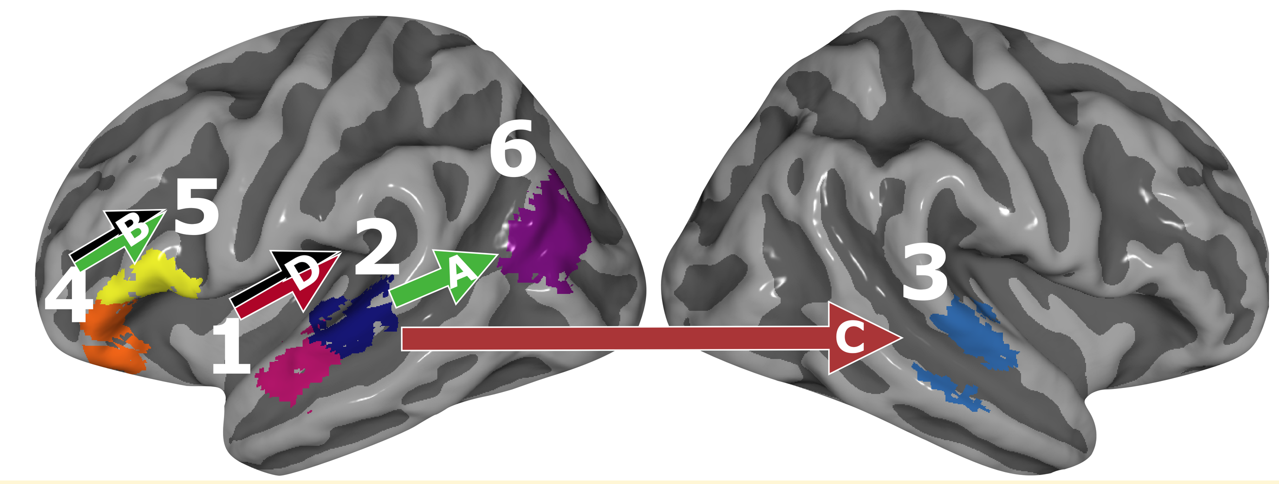


Encoding model performance significant in 34 language regions



Each of the proposed metrics reveals **distinct zone clusters**, that are consistent across datasets

Examples of the 4 types of inferences



Stimulus properties affect brain zones:

- mostly differently **(Inference A)**
- similarly & differently ELMo is missing properties that affect zones similarly **(Inference B)**
- mostly similarly **(Inference C)**
- similarly & differently ELMo is missing properties that affect zones differently **(Inference D)**

References

[1] Sebastian Weichwald, Timm Meyer, Ozan Özdenizci, Bernhard Schölkopf, Tonio Ball, and Moritz Grosse-Wentrup. Causal interpretation rules for encoding and decoding models in neuroimaging. Neuroimage, 110:48–59, 2015.

[2] Kendrick N. Kay, Thomas Naselaris, Ryan J. Prenger, and Jack L. Gallant. Identifying natural images from human brain activity. Nature, 452(7185):352, 2008.

[3] Shinji Nishimoto, An T. Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L. Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. Current Biology, 2011.

[4] Jean-Rémi King and Stanislas Dehaene. Characterizing the dynamics of mental representations: the temporal generalization method. Trends in cognitive sciences, 18(4):203–210, 2014.

[5] Mariya Toneva, Tom M. Mitchell, and Leila Wehbe. Combining computational controls with natural text reveals new aspects of meaning composition. bioRxiv, 2020.0

[6] Uri Hasson, Yuval Nir, Ifat Levy, Galit Fuhrmann, and Rafael Malach. Intersubject Synchronization of Cortical Activity during Natural Vision. Science, 303(5664):1634–1640, 3 2004.

[7] Boyle et al., The courtois project on neuronal modelling - 2021 data release. In Annual Meeting of the Organization for Human Brain Mapping, 2021.

[8] David C. Van Essen, Stephen M. Smith, Deanna M. Barch, Timothy E.J. Behrens, Essa Yacoub, and Kamil Ugurbil. The WU-Minn Human Connectome Project: An overview.NeuroImage, 80:62–79, 10 2013.