Proceedings of Machine Learning Research vol 140:1–27, 2022 1st Conference on Causal Learning and Reasoning

Same Cause; Different Effects in the Brain

Author names withheld

Editor: Under Review for CLeaR 2022

Abstract

To study information processing in the brain, neuroscientists manipulate experimental stimuli while recording participant brain activity. They can then use encoding models to find out which brain "zone" (e.g. which region of interest, volume pixel or electrophysiology sensor) is predicted from the stimulus properties. Given the assumptions underlying this setup, when the stimulus properties are predictive of the activity in a zone, these properties are understood to cause activity in that zone.

In recent years, researchers have begun using neural networks to construct representations that capture the diverse properties of complex stimuli, such as natural language or natural images. Encoding models built using these high-dimensional representations are often able to accurately predict the activity in large swathes of cortex, suggesting that the activity in all these brain zones is caused by the stimulus properties captured in the neural network representation. It is then natural to ask: "Is the activity in these different brain zones caused by the stimulus properties in the same way?" In neuroscientific terms, this corresponds to asking if these different zones process the stimulus properties in the same way.

Here, we propose a new framework that enables researchers to ask if the properties of a stimulus affects two brain zones in the same way. We use simulated data and two real fMRI datasets with complex naturalistic stimuli to show that our framework enables us to make such inferences. Our inferences are strikingly consistent between the two datasets, indicating that the proposed framework is a promising new tool for neuroscientists to understand how information is processed in the brain.

Keywords: encoding models, interpretability, neuroscience, fMRI, regression, neurolinguistics

1. Introduction

A major goal of neuroscience research is to understand how the brain processes information. To work towards this goal, neuroscientists need to map where information is processed in the brain. This type of mapping is frequently learned from encoding models which predict brain measurements from the properties of a stimulus (Mitchell et al., 2008; Kay et al., 2008; Nishimoto et al., 2011; Huth et al., 2012; Wehbe et al., 2014a; Huth et al., 2016; Schrimpf et al., 2020). Since encoding models can be built to predict measurements at different scales (e.g. functional Magnetic Resonance Imagining (fMRI) volume-pixel (voxel) or region of interest (ROI), electrophysiology sensor, etc.) we use the term brain "zone" to agnostically refer to measurement locations at any of these scales.

In a passive experimental paradigm (such as movie or image viewing) the stimuli are chosen by the experimenter and precede the brain measurements. With such experiments, we use encoding models to identify the brain zones in which the activity is associated with the stimulus properties. We also assume that the activity that is predicted using the encoding model is not related to some external factors or artifacts that are correlated with the stimulus. Under these assumptions, stimulus-based encoding models can be causally interpreted as revealing which brain zones are affected by the stimulus properties (Weichwald et al., 2015). It is worth it to note that in this stimulus-based setup, we intervene on the stimulus and control what information is presented to the participants. However, as we do not intervene directly on the brain (e.g., through transcranial magnetic stimulation (TMS)

© 2022 .

or electrical brain stimulation (EBS)) we cannot ask other, sometimes more crucial questions, such as whether the activity in a brain zone is *essential* for processing a stimulus.

If a stimulus-based encoding model tells us that two zones' measurements are affected by the stimulus properties, we can further increase our understanding of how information is processed in these zones by asking: do the stimulus properties affect both zones in the same way? This can be interpreted as asking if different subsets of the properties of the stimulus (e.g., in a movie stimulus, the language information and visual information) cause different effects in the two zones and/or if the same aspect of the stimulus (e.g. visual information) causes different effects in the two zones. As an illustration, consider two types of cells in the retina: on-center and off-center cells. Consider two such cells that process input from the same visual area. Flashing a small light in the center of this shared receptive field will affect the response of both cells. However, the on-center cell's firing rate will increase, while the off-center cell will be inhibited (Kuffler, 1953). The same cause leads to two different effects, and distinguishing these effects is key to understanding the function of the cells.

In settings with more complex stimuli (including naturalistic stimuli such as watching movies, reading books, listening to stories) it becomes hard to infer if the stimulus affects different zones in the same way as there are multivariate and often high-dimensional aspects to the stimulus. In fact this difficulty is why it remains an open question if ROIs in the language network (Fedorenko et al., 2010) which are affected by language stimuli, are affected in the same way (Wehbe et al., 2014a; Reddy and Wehbe, 2020; Caucheteux et al., 2021a; Huth et al., 2016; Deniz et al., 2019). Given that it is becoming increasingly popular in neuroscience to use more complex naturalistic stimuli that are more faithful to the real world (Sonkusare et al., 2019; Nastase et al., 2020; Hamilton and Huth, 2020), it is important to enable such inferences in settings with complex stimuli.

An encoding model relies on a representation of the stimulus that corresponds to a specific hypothesis (e.g. a syntactic feature space can represent the linguistic structure of sentences and be used to identify brain zones processing syntax). Here, we use the term *stimulus properties* to refer to the real, latent features of the stimulus that affect brain activity, and the term *stimulus-representation* to refer to the vector representation that the experimenter builds to approximate the stimulus properties. The choice of the stimulus-representation is important, and will have consequences on the modeling and the conclusions. It is increasingly popular to use large neural networks as models for the information represented in the brain, due to their rich expressivity (Yamins et al., 2014; Wehbe et al., 2014b; Jain and Huth, 2018; Toneva and Wehbe, 2019; Wang et al., 2019; Schrimpf et al., 2020; Caucheteux et al., 2021b; Goldstein et al., 2021; Cross et al., 2021). However, because the representations in these networks are high-dimensional and combine different types of information about the stimulus, they often predict a large portion of cortex. This raises the need for more refined approaches that not only detect a causal effect between the stimulus properties and the brain activity, but that also can determine whether the effect is similar or different across regions. Generally, this setting fits under the umbrella of outcome-wide design (VanderWeele, 2017).

In this work, we present a new framework that includes two new metrics which enable researchers to move beyond asking whether the activity in a zone is affected by a stimulus to asking if this stimulus affects different zones in the same way. The first metric captures how much two zones respond in the same way to the aspects of a stimulus captured by a *stimulus-representation*. We term this first metric **zone generalization**. The second metric captures the magnitude of any *stimulus*-related response that is not shared between the two zones (even if this stimulus-related response is created by information not included in the stimulus representation). We term this second metric **zone residuals**. We introduce an implementation of both types of metrics in Section 2.2 and 2.3.

We first present our framework as a decision tree in Fig. 1, which researchers can follow to infer if a stimulus affects two brain zones in the same way. In this work, we employ linear functions to relate the metrics in the framework to the most frequently used type of encoding model—the linear encoding model. However our framework is general as the metrics can be defined according to any predictive function. Next, we present simulations that show that (1) when used together zone generalization and zone residuals enable researchers to infer if a stimulus affects two brain zones with significant encoding model performance in the same way and that (2) these two new metrics are needed because the commonly used methods for studying information processing in the brain cannot be used to make this inference. Importantly, we make our simulation data and code available so that researchers can test whether other implementations of these two metrics enable this inference (with their preferred statistical models). We showcase the use of our framework on two fMRI datasets with complex naturalistic stimuli, and show when we can infer if a stimulus affects two brain zones in the same way and when this inference cannot be made. Our inferences generalize across these two datasets which capture different populations of participants and were acquired by different labs, in different countries, with different movie stimuli and different scanning parameters. This generalizability highlights the robustness of our approach.

Framework as a Decision Tree: A researcher creates a stimulus-based encoding model to identify brain zones that are well predicted by a given stimulus-representation (e.g., ELMo (Peters et al., 2018)). This can be measured with encoding model performance. Assume that this approach reveals that the activity in two different brain zones is predicted by ELMo. The typical encoding model pipeline stops analyzing how information is processed in the brain here.

In our framework (Fig. 1), the researcher instead proceeds to computing zone generalization, measuring how much two zones respond in the same way to given aspects of a stimulus.

1. If zone generalization is large, we conclude the brain zones respond similarly to the stimulus properties captured by the stimulus-representation. It is still unclear however if the zones respond similarly to the entire stimulus, or if the stimulus-representation is incomplete and does not include stimulus properties that affect the zones differently. Here, we use our second metric, zone residuals, that captures the magnitude of any stimulus-related response that is not shared between the two zones, even outside of the stimulus-representation:
   (a) If (normalized) zone residuals are small, we conclude that the two zones are responding mostly similarly to the stimulus, and that the stimulus properties the zones respond similarly to are (at least in part) captured by the stimulus-representation.
   (b) If (normalized) zone residuals are large, we conclude that the two zones are responding similarly to some properties and differently to others, however the stimulus-representation is incomplete and does not capture stimulus properties that affect zones differently.
2. If zone generalization is small, we conclude that the two zones respond differently to the stimulus properties captured in the stimulus-representation. It is still unclear if the zones are responding differently to the entire stimulus, or if the stimulus-representation is incomplete and does not include all the stimulus properties that affect zones similarly. Here, we use our second metric again:
   (a) If (normalized) zone residuals are large, then we conclude that the two zones are responding mostly differently to the stimulus.
   (b) If (normalized) zone residuals are small, then we conclude that the two zones are responding similarly to some stimulus properties and differently to others, however the stimulus-representation is incomplete and does not capture all the stimulus properties that affect zones similarly.

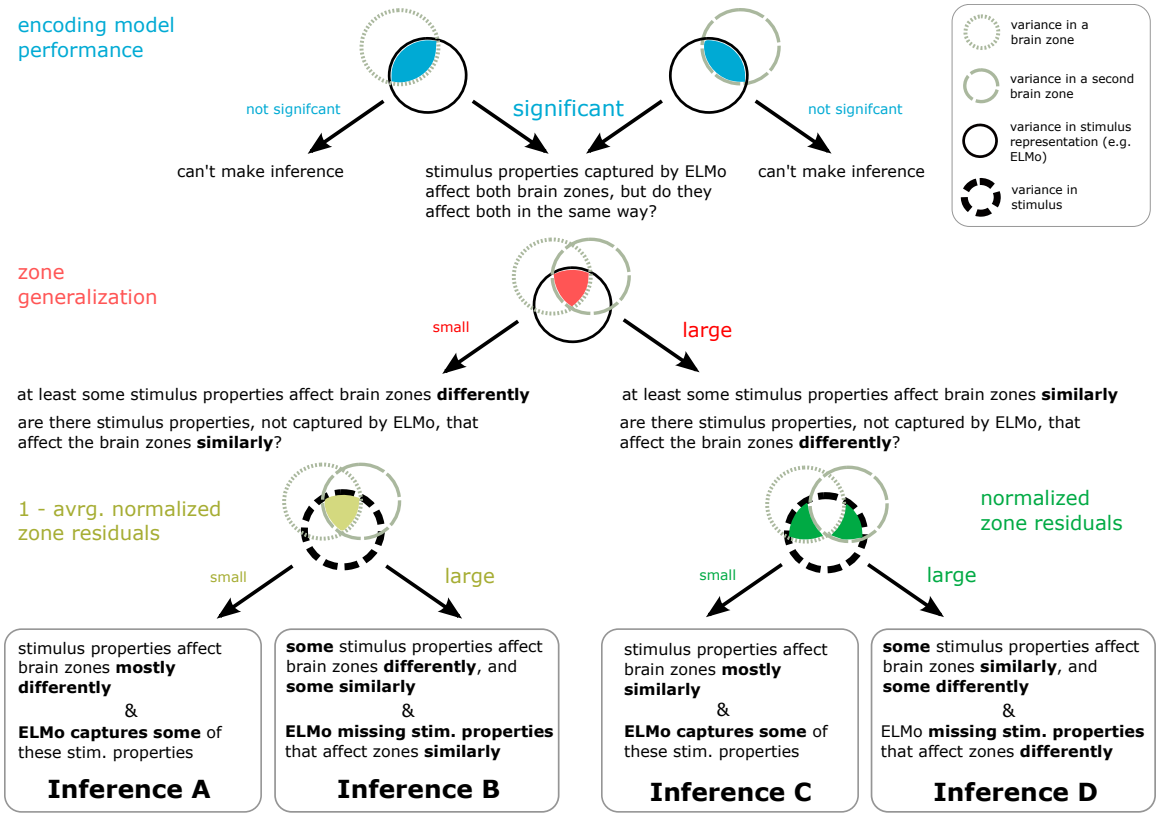


Figure 1: Proposed framework to infer whether two brain zones are affected by stimulus properties in the same way. The first step follows current work and computes the encoding model performance of predicting each of the brain zones as a function of a stimulus-representation (i.e. a representation from a deep neural network, such as ELMo (Peters et al., 2018) for language stimuli). We cannot infer directly from the encoding model performance whether the two brain zones are similarly affected by the stimulus, so we proceed to the second step. Large zone generalization allows us to infer that the brain zones respond similarly to the stimulus properties captured by the stimulus-representation. At the last step, normalized source residuals allow us to infer the magnitude of any stimulus-related response that is not shared between the two zones, even outside of the stimulus-representation. We use both metrics to make one of four inferences.

2. Definitions

To ground the definitions of our metrics, we introduce three settings of how brain measurements could be generated that are increasingly more specific about the dependence between one brain zone and other zones, in the same participant and other participants. All settings assume that all brain zones are normalized to have mean 0 and variance 1, which is a standard preprocessing step.

2.1. Single brain zone, single participant

In the first setting, the activity in a single brain zone that is recorded from one participant is assumed to be generated as a function of a specific representation of the presented stimulus. More concretely,

$$Y_i = g_i(X) + \epsilon \tag{1}$$

where $Y_i \in \mathbb{R}$ is the observation at a brain zone i (e.g. fMRI voxel or ROI, EEG/MEG sensor, etc.), $X \in \mathbb{R}^d$ is the d-dimensional numerical representation of the corresponding stimulus (e.g. a word embedding obtained from inputting a word stimulus in a language model), and $\epsilon \in \mathbb{R}$ to a noise term that is independent from the stimulus-representations.

Encoding model performance. This first setting is the one that is assumed by encoding models (Mitchell et al., 2008; Kay et al., 2008; Nishimoto et al., 2011; Huth et al., 2012; Wehbe et al., 2014a; Huth et al., 2016). An encoding model estimates for each zone i the function $g_i(\cdot)$ for a specific stimulus-representation X. Most commonly, this function is parameterized as a linear function (i.e. $g_i(X) = \langle X, \theta_i \rangle$, where $\theta_i \in \mathbb{R}^d$), and is estimated using a set of training observations for each zone in a cross-validated fashion. The performance of an encoding model in zone i is often evaluated as the Pearson correlation of held-out data Y_i and the corresponding predictions $\widehat{Y}_i = \hat{g}_i(X)$:

$$\texttt{encoding model performance}\,(\text{zone}_i) = \text{corr}(\widehat{Y}_i, Y_i).$$

2.2. Multiple brain zones, single participant

In the second setting, we explicitly allow for a noise term reflecting shared information between two zones Y_1 and Y_2 that is not driven by the stimulus and not captured by the stimulus-representation:

$$Y_i = \underbrace{g_i(X)}_{\text{signal}} + \underbrace{\epsilon_i}_{\text{individual noise}} + \underbrace{\epsilon_{12}}_{\text{shared noise}}, \qquad i \in \{1, 2\} \tag{2}$$

where $g_i(X) = \langle X, \theta_i \rangle$ are linear functions of stimulus-representations $X \in \mathbb{R}^d$ with parameters $\theta_i \in \mathbb{R}^d$ and $\epsilon_1, \epsilon_2, \epsilon_{12}$ are independent noise terms with zero mean and variances σ_1^2, σ_2^2 and σ_{12}^2.

First metric in framework: zone generalization. Our goal is to infer whether a stimulus affects two zones in the same way. As outlined in Section 1, we are not able to rely on encoding model performance for this inference. To address this limitation, we propose *zone generalization*, which estimates the amount of information shared by two zones and captured in the stimulus-representation:

$$\texttt{zone generalization}\,(\text{zone}_1, \text{zone}_2) = \text{corr}(\widehat{Y}_1, Y_2)\,. \tag{3}$$

$\widehat{Y}_1 = \hat{g}_1(X)$ is the prediction from an encoding model trained on zone 1's data (with X corresponding to held-out stimuli). Y_2 is zone 2's data recorded when the same held-out stimuli was presented. Note that while zone generalization can be estimated in other ways, we chose this specific implementation because of its similarity to and ease of comparison with encoding model performance.

Large zone generalization indicates that zone 1 and 2 respond in the same way to at least some stimulus properties. Small zone generalization, when accompanied by a significant encoding performance for both zones, indicates that zone 1 and 2 respond differently to at least some stimulus properties. However, these possible inferences are fundamentally limited by the quality of the stimulus-representation and how much it captures stimulus properties. For example, if zone 1 and 2 are affected by some stimulus properties in the same way, but the stimulus-representation does not include these properties, zone generalization may be small.

2.3. Multiple brain zones, multiple participants

Our third setting allows for multiple participants $P \in \{1, 2, ...M\}$. The data for each participant is standardized to the same anatomical space, and each zone i is defined for each participant. Each quantity in Equation (2) is allowed to differ in each participant P:

$$Y_{i,P} = \underbrace{g_{i,P}(X)}_{\text{signal of stimulus-representation}} + \underbrace{h_{i,P}(Z)}_{\text{signal of } Z} + \underbrace{\epsilon_{i,P}}_{\text{individual noise}} + \underbrace{\epsilon_{12,P}}_{\text{joint noise}}, \qquad i \in \{1, 2\}. \tag{4}$$

Additionally we include a $h_{i,P}(Z)$ component which makes explicit a dependence on features Z that are part of the stimulus, but are not captured by the stimulus-representation X.

Second metric in framework: zone residuals. Zone generalization enables us to infer whether two zones respond similarly or differently to at least some stimulus properties, but it may be unable to present the complete picture in the case that the stimulus-representation does not capture all stimulus properties. To address this limitation, we introduce *zone residuals*, which capture stimulus-related variance that is unique to a brain zone. Our implementation of the metric builds on the intuition behind inter-subject correlation (Hasson et al., 2004) to estimate how much of the activity in zone 1 (or 2) is affected by the stimulus but not shared with zone 2 (or 1). More concretely, zone residuals are estimated between each pair of participants S and T and then averaged across all pairs:

$$\texttt{zone residuals}(\text{zone}_1, \text{zone}_2) = \frac{1}{M^2 - M} \sum_{S,T,S \neq T} \text{corr}(R_{1-2,S}, R_{1-2,T}), \tag{5}$$

where $R_{1-2,P} = Y_{1,P} - \text{corr}(Y_{1,P}, Y_{2,P})Y_{2,P}$ is the residual of regressing $Y_{2,P}$ from $Y_{1,P}$. The residual $R_{1-2,P}$ is obtained from the activity in zone 1 that cannot be predicted by the activity in zone 2. Correlating these residuals across participants allows us to estimate how much of this activity is consistent across participants, i.e. driven by the stimulus. The intuition is that significant similarity between the brain activity of two people viewing the same stimulus must be driven by that stimulus (Hasson et al., 2004; Hebart et al., 2018). Thus, correlating the residuals that we obtain from the brain zone-to-brain zone prediction allows us to identify whether there is any unique activity in one zone that is stimulus-dependent. When paired with encoding model performance and zone generalization, zone residuals can help infer whether stimulus properties affect two brain zones similarly or differently, or whether the brain zones are affected similarly by some stimulus properties and differently by others.

2.4. Metric Normalization

Many of the metrics discussed in the previous subsections may be more interpretable when they are related to the amount of predictable brain activity (i.e. a "noise ceiling"). The noise ceiling we use here is the square root of inter-subject correlation, similarly to Wehbe et al. (2021). Inter-subject correlation reveals which zones are correlated across participants as they process the same stimulus, thus revealing zones that respond to the stimulus (Hasson et al., 2004). For each zone i:

$$\texttt{inter-subject correlation}\,(\text{zone}_i) = \frac{1}{M^2 - M} \sum_{S,T,S \neq T} \text{corr}(Y_{i,S}, Y_{i,T})\,. \tag{6}$$

Whether the raw values for the individual metrics should be used or whether they should be put in perspective with a noise ceiling or a different quantity depends on the question under investigation. We refer the reader to Appendix A, where we include a figure and discussion to expand this point.

3. Simulations

Here, we examine if our proposed framework can help us infer if a stimulus affects brain zones in the same way. Working with synthetic data where we know the ground truth helps provide intuition for how informative each metric discussed in Section 2 can be in a variety of controlled scenarios.

Generating synthetic data. Consider an experiment where two participants are presented with the same naturalistic stimulus while their brain activity is being recorded at the same two brain zones. Suppose also that we decide on some numerical representation of the presented stimulus. To generate synthetic data, we make the simplifying assumption that all stimulus properties are captured by one of the following two disjoint representations: (i) $X \in \mathbb{R}^d$, the stimulus-representation itself or (ii) $Z \in \mathbb{R}^d$, a representation that captures all remaining stimulus properties that X does not capture. We then synthesize brain activity measurements at each zone as a linear function of X, Z and an additional term ϵ that represents stimulus-independent noise observed in the brain zone. Our data generation model includes mechanisms that allow us to vary how much of the stimulus-driven activity is due to both (1) the two brain zones responding similarly to the stimulus properties (vs. differently) and (2) stimulus properties captured by the stimulus-representations (vs. missing from). Details on these mechanisms and the overall data generation model are provided in Appendix B.

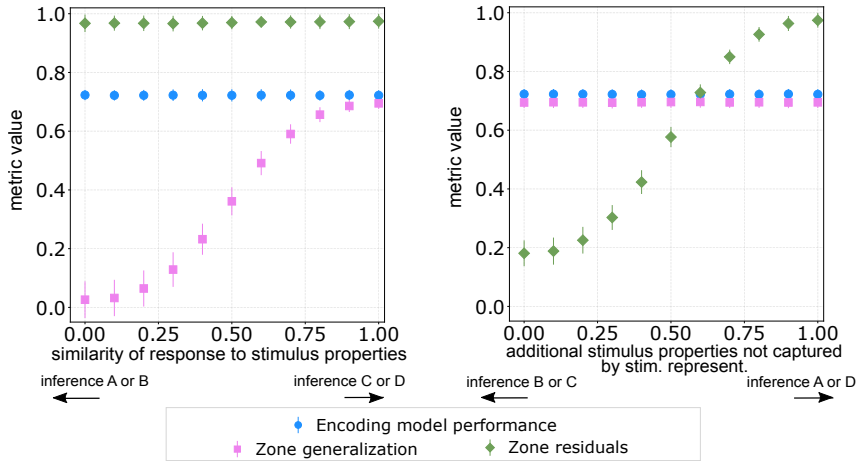


Figure 2: Plotting average metric values under simulations that separate (Left) inference A or B from inference C or D and (Right) inference B or C from inference A or D.

Separating inferences A, B from inferences C, D. In the decision tree in Fig. 1, inferences A and B are separated from inferences C and D based on whether the two zones respond similarly to the stimulus properties captured in the stimulus-representation. We vary how similarly the two zones respond to the stimulus properties and observe the effect on our metrics. Fig. 2(Left) shows how each metric, computed and averaged over 1000 repetitions, varies as we do so. We find that encoding model performance and zone residuals remain relatively unchanged across the board. Only zone generalization seems to be informative of how similar two brain zones respond to stimulus properties, to help determine which pair of inferences (A or B vs. C or D) can be made.

Separating inference A or D from inference B or C. From Fig. 1, one can separate inference A from B, and inference C from D, if one has information about the extent to which both zones respond to stimulus properties not captured by the stimulus-representation. In Fig. 2(Right) we plot metric values, computed and averaged across 1000 repetitions, as we vary the proportion of stimulus properties that are driving the brain zones but are not captured by the stimulus-representation. We find that of the three metrics of interest, only the zone residuals varies when this proportion is

changed. This is because increasing the proportion of additional stimulus properties also increases the proportion of stimulus properties that affect the zones differently, leading to larger zone residuals. Thus, after we use zone generalizations to narrow our search down to a pair of inferences (A or B vs. C or D), zone residuals can be used to precisely infer how the stimulus affects the pair of zones.

An important caveat to note is that zone generalization itself depends on the goodness of fit of the encoding model for the original brain zone. If the pair of zones we are investigating both have significant encoding model performances, zone generalization can be informative for inferring if the zones respond similarly to the stimulus properties captured by the stimulus-representation. But if the encoding model performance is non-significant for one or both zones, then we cannot even infer if the stimulus properties captured by the stimulus-representation affect both zones. Non-significant encoding model performance could be due to low SNR or an incomplete stimulus-representation that does not capture any of the stimulus properties that the zone(s) respond to.

Lastly, note that the halfway point between inferences A/B and inferences C/D occurs near 0.4 zone generalization (Fig. 2(Left)), and the halfway point between inferences B/C and inferences A/D occurs near 0.6 zone residual (Fig. 2(Right)). In the real fMRI data experiments in the next section, we use these values as the decision boundaries for the inferences depicted in Fig. 1.

4. Empirical Results on Two Naturalistic fMRI Datasets

Here, we examine if our proposed framework can help us infer if a stimulus affects different brain zones in the same way. We use two fMRI datasets with complex naturalistic stimuli (i.e., video clips), specifically because they present a trade-off between the number of participants and the amount of data recorded for each participant, enabling us to evaluate our framework under different constraints.

HCP: short movies. We use publicly available data from the Human Connectome Project (HCP) 7T dataset, with healthy participants between 22-36 years old (Van Essen et al., 2013). HCP fMRI data comes minimally pre-processed as FIX-Denoised data (Glasser et al., 2013; Griffanti et al., 2014; Salimi-Khorshidi et al., 2014). We analyze data from 90 participants that were randomly selected from the full dataset for exploratory purposes. Each participant watched naturalistic audio-visual video clips in English. In total, 60 minutes and 55 seconds of data were recorded per participant during 4 scans of approximately equal length. The fMRI sampling rate (TR) was 1 second.

Courtois NeuroMod data: full-length movie. The second fMRI dataset is provided by the Courtois NeuroMod group (data release `cneuromod-2020`). In this dataset, 6 healthy participants view the movie Hidden Figures in English. In total, approximately 120 minutes of data were recorded per participant during 12 scans of roughly equal length. The fMRI sampling rate (TR) was 1.49 seconds. This data is available by request at `https://docs.cneuromod.ca/en/latest/ACCESS.html`.

Other data processing details. For each participant we downsample the fMRI data by averaging the voxel activities within the 268 functionally defined ROIs from the Shen atlas (Shen et al., 2013; Finn et al., 2015), similarly to previous work (Rosenberg et al., 2015; Greene et al., 2018; Gao et al., 2019; Doss et al., 2020). For each participant this results in a dataset of dimensions number of TRs $\times$ 268 ROIs. These ROIs are entirely independent of our data as the Shen atlas was previously constructed from a separate group of healthy participants. Because we are interested in studying naturalistic language comprehension, we chose to identify Shen atlas ROIs involved with processing language-relevant information. We use language ROIs identified by Fedorenko et al. (2010) and semantic ROIs (Binder et al., 2009) (also defined entirely independently of our data). We consider a Shen atlas ROI to be a language ROI if > 15% of its voxels are within one of those previously identified ROIs. This procedure selects 55 Shen atlas ROIs (more details in Appendix C).

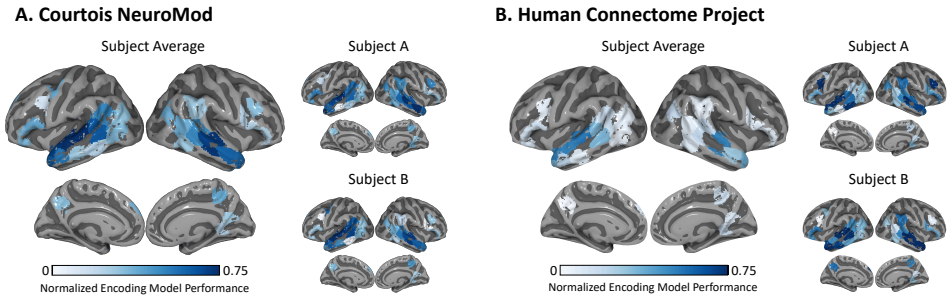


Figure 3: Encoding model performance at 34 significantly predicted ROIs (corrected at level 0.05).

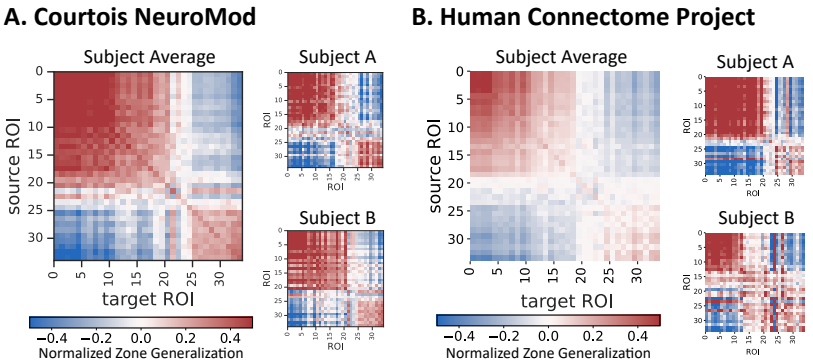
Stimulus-representation. The proposed framework is general and can be applied to a wide variety of stimulus-representations. Because we are specifically interested in studying the processing of language-relevant information, we follow previous work and obtain representations of the stimulus words by feeding the transcripts of the speech in the videos word-by-word into a pre-trained natural language processing model. We choose ELMo (Peters et al., 2018), a bidirectional language model with multiple LSTM layers. Word representations obtained from the first hidden layer of ELMo, and contextualized with the previous 25 words, have been shown to significantly predict fMRI recordings of participants comprehending language (Toneva and Wehbe, 2019; Toneva et al., 2020). We focus our analyses on representations similarly collected from the first hidden layer of pre-trained ELMo (Gardner et al., 2017) when provided with chunks of 25 consecutive words.

Encoding model performance. Firstly, the framework computes encoding model performance, which reveals which brain zones are affected by the stimulus properties captured by ELMo. We estimate encoding models which predict the brain activity associated with watching video clips from the ELMo embedding of the speech in the video clip. One encoding model is estimated independently for each ROI in each participant. The model is parameterized as a linear function regularized by the ridge penalty similarly to previous work (Sudre et al., 2012; Wehbe et al., 2014b,a; Nishimoto et al., 2011; Huth et al., 2016; Toneva and Wehbe, 2019) and evaluated using cross-validation (CV), where one of the scans is heldout for testing during each CV fold (i.e. 4 folds for HCP, and 12 folds for Courtois NeuroMod). The regularization parameter is chosen by nested 10-fold CV.

We identify 34 bilateral language ROIs that are significantly predicted across participants in both fMRI datasets (one-sample t-test, FDR corrected for multiple comparisons across ROI at alpha level 0.05 (Benjamini and Hochberg, 1995)), suggesting that these ROIs are affected by the stimulus properties captured in ELMo. These results replicate previous findings that the language ROI are well predicted by representations from ELMo (Toneva and Wehbe, 2019; Toneva et al., 2020). For these 34 ROIs we present the average normalized encoding performance across all participants in the datasets, as well as the normalized encoding performance for two representative participants in Fig. 3 (see Appendix Fig. 13 for additional participant-level performances). The encoding model performances are normalized by inter-subject correlation as described in section 2.4.

Zone generalization. To answer if the stimulus properties captured by ELMo affect the 34 language ROIs in the same way, we turn to the next step of the framework and compute zone generalization. We present pairwise normalized zone generalization for the 34 language ROIs in both fMRI datasets in Fig. 4 (see Appendix Fig. 16 for the participant-level zone generalization). To visualize where the ROIs associated with each ROI pair are located on the brain see the Appendix Fig. 11. The zone generalization values are also normalized by the inter-subject correlation. In both datasets, we find that there are large normalized zone generalization values (shown in red) which

Figure 4: Zone Generalization. ROI pairs with large norm. zone generalization (red) are affected similarly by at least some stimulus properties. Pairs with large norm. zone generalization are consistent at the group and participant-level in both datasets.



means at least some stimulus properties affect those ROI pairs similarly. These large normalized zone generalization values are consistent at the group and participant-level in both datasets. In both datasets, we also find other pairs of ROI that exhibit small normalized zone generalization values, including negative values (see Appendix E for additional discussion about negative normalized zone generalization), which means that at least some stimulus properties affect those ROI pairs differently.

Zone residuals. At this point in the framework, it is unclear if ELMo is incomplete and does not capture relevant stimulus properties. We turn to the last step and compute the zone residuals. We present the normalized zone residuals for the 34 language ROIs in both datasets in Appendix Fig. 12 (see Appendix Fig. 15 for participant-level normalized zone residuals). The zone residuals are normalized by the inter-subject correlation. In both datasets, we find large (dark green) and small normalized zone residuals at the group and participant-level. The ROI pairs with large normalized zone residuals are consistent at the group and participant-level in both datasets.

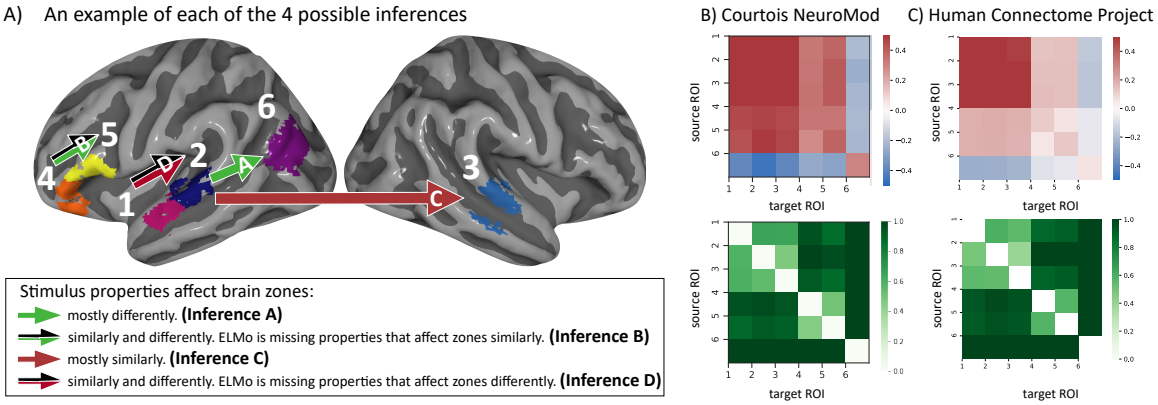


Figure 5: We use the proposed framework to infer answers to the question: Does the stimulus affect both brain zones the same way? We present an example of each of the four possible inferences.

When to make each of the four inferences. To make inference A-D using the framework we need to use encoding model performance, zone generalization and zone residuals. As we aim to study naturalistic language comprehension, and to infer if the stimulus affects pairs of ROIs in the same way, we focus on six language ROIs in which it has been previously shown it is difficult to make this inference (Reddy and Wehbe, 2020; Caucheteux et al., 2021a). We show an example of each of the four possible inferences that can be made using the proposed framework within the six ROIs in Fig. 5. These relationships are consistent across both datasets (see Appendix Figs. 16, 17 for

participant-level heatmaps). The relationships between ROI pairs can be asymmetric (ie. ROI i could better generalize to ROI j than ROI j to ROI i), therefore we depict relationships as a directed edge from a source to a target ROI. Inference A, between ROI 2 (left posterior temporal gyrus) and ROI 6 (left angular gyrus) indicates that while they are both significantly predicted by ELMo, they respond differently to language properties captured by ELMo. Inference B between ROI 4 (left inferior frontal gyrus, pars orbitalis) and ROI 5 (left inferior frontal gyrus) indicates that both ROIs similarly process some language properties that are not included in ELMo. Inference D between ROI 1 and ROI 2 (anterior and posterior left temporal gyri) indicates that while these ROIs do process some properties similarly, ELMo is still missing some language properties that they process differently. Finally, Inference C between ROI 2 and ROI 3 (left and right posterior temporal gyri) supports the hypothesis that some language properties are processed similarly in bilateral posterior temporal cortex.

Discussion. In contrast to encoding model performance alone, our proposed framework can be used to make inferences A-D as outlined in Fig.1. Note that in inference B both zone generalization and zone residuals are small, and in inference D both zone generalization and zone residuals are large. In both cases, the stimulus-representation does not capture the relevant stimulus properties and our modeling can be improved by adding a stimulus-representation that does. For example, our results suggests that ROI 1 and ROI 2 (anterior and posterior left temporal gyri) in Fig. 5 may be easier to distinguish if we add stimulus-representations that capture the unique properties that either ROI responds to. Our framework can be used as a test for future stimulus-representations—if these stimulus-representations lead to significantly larger encoding model performance for either ROI 1 or ROI 2, while the zone generalization and zone residuals do not change significantly, then the new stimulus-representation better captures some of the unique properties that affect at least one ROI.

5. Related Work

To study information processing in the brain, researchers typically investigate either the relationship between brain zones, or the relationship between a brain zone and a stimulus. The most common approach to study the relationship between brain zones is functional connectivity (Friston et al., 1993). Most frequently, functional connectivity correlates the measurements between two zones (Mohanty et al., 2020) of the same participant. Obtaining the functional connectivity between zone 1 and zone 2 in our setting is equivalent to computing $\text{corr}(Y_1, Y_2)$. Outside of encoding models and inter-subject correlation, a popular approach to relate brain zones and stimuli is representational similarity analysis (RSA) (Kriegeskorte et al., 2008) (which can also be used to relate two brain zones). RSA measures the similarity of two representational dissimilarity matrices, that each describe the distances between pairs of stimuli in one realm (e.g. a stimulus-representation or a brain zone).

As brain measurements contain stimulus related activity, non-stimulus related activity, and noise, the similarity between the measurements of two zones does not provide specific insights into the relationship between the zones and the stimulus. We show in the simulation in Appendix B.4 that both functional connectivity and $\text{RSA}(\text{zone}_1, \text{zone}_2)$ do not change when we vary the similarity between the two zones, or the proportion of stimulus properties that are in the stimulus-representation. Thus, neither metric can help us distinguish between inferences A-D. Further, $\text{RSA}(\text{zone}_1, \text{stimulus-representation})$ and $\text{RSA}(\text{zone}_2, \text{stimulus-representation})$ also do not vary when we change the simulation settings. This is not surprising since RSA between a brain zone and a stimulus-representation bears similarity with encoding model performance—they both estimate the strength of the relationship between the brain zone and the stimulus-representation, though the RSA value is less readily interpretable. Like encoding model performance, RSA does not

help distinguish between the four inferences. A more complicated RSA setup could be devised in which the RSA is computed between predictions $\hat{Y}_1$ and real data Y_2, or between the (zone 1 and zone 2) residuals of two participants. We consider such setups to be alternative implementations of our zone generalization and zone residuals metrics.

To investigate if brain zones respond similarly or differently to the stimulus, researchers often use encoding model weights (Çukur et al., 2013; Huth et al., 2012, 2016; Deniz et al., 2019). Encoding model weights can reveal the latent directions of largest variance in the tuning to a stimulus-representation (Huth et al., 2012, 2016; Deniz et al., 2019) and even how tuning in the same zone is affected by a specific task (i.e., attending to humans) (Çukur et al., 2013). The latent directions of largest variance can be revealed by obtaining the principal components of the brain zone weights and plotting the first few principal component scores on the brain (Huth et al., 2012, 2016; Deniz et al., 2019). The change of tuning in the same zone due to attention can be revealed by correlating a zone's encoding model weights with a binary template of the attended stimulus (Çukur et al., 2013). These approaches thus evaluate changes in tuning qualitatively (by producing interpretable brain maps) and quantitatively (by computing metrics on the weights).

One can think of zone generalization as a quantitative tool to estimate tuning differences not through the weights but through the predictions of new data. Zone generalization is a general metric that builds upon temporal generalization (King and Dehaene, 2014) and spatial generalization (Toneva et al., 2020), which can be considered two specific instances of this metric. King and Dehaene (2014) proposed temporal generalization, to allow researchers to see if the pattern of responses to a stimulus at a given point in time are similar to another point in time, subsequently this has been used in multiple works (Blanco-Elorrieta and Pylkkänen, 2017; Hebart et al., 2018; Fyshe et al., 2019; Fyshe, 2020). Then Toneva et al. (2020) adapted this temporal generalization idea to compare the pattern of responses to a stimulus in different voxels. Spatial generalization uses the predictions of one voxel's measurements to predict the measurements in another voxel. To the best of our knowledge we are not aware of any existing metrics that are similar to zone residuals.

6. Conclusion and Future Work

We presented a new framework including two metrics, zone generalization and zone residuals, to enable researchers to infer if a stimulus affects brain zones in the same way. We showed in simulation that when used in addition to significant encoding model performance, zone generalization and zone residuals enable this inference, while commonly used methods for studying information processing do not. Finally, we showed that the results from our framework generalize across two naturalistic fMRI datasets which capture different populations and are acquired by different labs in different countries with very different experimental setups and scanning parameters.

One limitation of the zone residuals metric is that when subtracting the prediction of one zone from the activity of another, the resulting residual contains information about the unique response of both zones. Isolating the residual information in an individual zone may further improve our ability to infer how different zones respond to stimulus properties. We hope that our approach can serve as basis for such future work. Additionally, to make inferences in the fMRI datasets, we base the thresholds for the decision points illustrated in Fig. 1 on the simulated data. While these thresholds lead to repeatable results across two real datasets, future work that establishes principled dataset-specific thresholds may be fruitful. We hope that our contributions can inspire such future work. Overall, our proposed framework is a promising new tool for computational neuroscientists who are interested in understanding how information is processed in the brain.

References

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Jeffrey R Binder, Rutvik H Desai, William W Graves, and Lisa L Conant. Where is the semantic system? a critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral cortex*, 19(12):2767–2796, 2009.

Esti Blanco-Elorrieta and Liina Pylkkänen. Bilingual language switching in the laboratory versus in the wild: The spatiotemporal dynamics of adaptive language control. *Journal of Neuroscience*, 37 (37):9022–9036, 2017.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Remi King. Decomposing lexical and compositional syntax and semantics with deep language models. *arXiv preprint arXiv:2103.01620*, 2021a.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Remi King. Gpt-2's activations predict the degree of semantic comprehension in the human brain. *bioRxiv*, 2021b.

Logan Cross, Jeff Cockburn, Yisong Yue, and John P O'Doherty. Using deep reinforcement learning to reveal how the brain encodes abstract state-space representations in high-dimensional environments. *Neuron*, 109(4):724–738, 2021.

Tolga Çukur, Shinji Nishimoto, Alexander G. Huth, and Jack L. Gallant. Attention during natural vision warps semantic representation across the human brain. *Nature Neuroscience*, 16(6):763–770, 6 2013. ISSN 10976256. doi: 10.1038/nn.3381. URL https://www.nature.com/articles/nn.3381.

Fatma Deniz, Anwar O Nunez-Elizalde, Alexander G Huth, and Jack L Gallant. The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality. *Journal of Neuroscience*, 39(39):7722–7736, 2019.

Manoj K. Doss, Darrick G. May, Matthew W. Johnson, John M. Clifton, Sidnee L. Hedrick, Thomas E. Prisinzano, Roland R. Griffiths, and Frederick S. Barrett. The Acute Effects of the Atypical Dissociative Hallucinogen Salvinorin A on Functional Connectivity in the Human Brain. *Scientific Reports*, 10(1):16392, 12 2020. ISSN 20452322. doi: 10.1038/s41598-020-73216-8. URL https://doi.org/10.1038/s41598-020-73216-8.

Oscar Esteban, Ross Blair, Christopher J. Markiewicz, Shoshana L. Berleant, Craig Moodie, Feilong Ma, Ayse Ilkay Isik, Asier Erramuzpe, Mathias Kent, James D. andGoncalves, Elizabeth DuPre, Kevin R. Sitek, Daniel E. P. Gomez, Daniel J. Lurie, Zhifang Ye, Russell A. Poldrack, and Krzysztof J. Gorgolewski. fmriprep. *Software*, 2018a. doi: 10.5281/zenodo.852659.

Oscar Esteban, Christopher Markiewicz, Ross W Blair, Craig Moodie, Ayse Ilkay Isik, Asier Erramuzpe Aliaga, James Kent, Mathias Goncalves, Elizabeth DuPre, Madeleine Snyder, Hiroyuki Oya, Satrajit Ghosh, Jessey Wright, Joke Durnez, Russell Poldrack, and Krzysztof Jacek Gorgolewski. fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, 2018b. doi: 10.1038/s41592-018-0235-4.

Evelina Fedorenko, Po-Jang Hsieh, Alfonso Nieto-Castañón, Susan Whitfield-Gabrieli, and Nancy Kanwisher. New method for fMRI investigations of language: defining ROIs functionally in individual subjects. *Journal of neurophysiology*, 104(2):1177–1194, 8 2010. ISSN 1522-1598. doi: 10.1152/jn.00032.2010. URL https://pubmed.ncbi.nlm.nih.gov/20410363https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2934923/.

Emily S. Finn, Xilin Shen, Dustin Scheinost, Monica D. Rosenberg, Jessica Huang, Marvin M. Chun, Xenophon Papademetris, and R. Todd Constable. Functional connectome fingerprinting: Identifying individuals using patterns of brain connectivity. *Nature Neuroscience*, 18(11):1664–1671, 11 2015. ISSN 15461726. doi: 10.1038/nn.4135. URL https://www.nature.com/articles/nn.4135.

K. J. Friston, C. D. Frith, P. F. Liddle, and R. S.J. Frackowiak. Functional connectivity: The principal-component analysis of large (PET) data sets. *Journal of Cerebral Blood Flow and Metabolism*, 13(1):5–14, 1 1993. ISSN 0271678X. doi: 10.1038/jcbfm.1993.4. URL http://journals.sagepub.com/doi/10.1038/jcbfm.1993.4.

Alona Fyshe. Studying language in context using the temporal generalization method. *Philosophical Transactions of the Royal Society B*, 375(1791):20180531, 2020.

Alona Fyshe, Gustavo Sudre, Leila Wehbe, Nicole Rafidi, and Tom M Mitchell. The lexical semantics of adjective–noun phrases in the human brain. *Human brain mapping*, 40(15):4457–4469, 2019.

James S. Gao, Alexander G. Huth, Mark D. Lescroart, and Jack L. Gallant. Pycortex: an interactive surface visualizer for fMRI. *Frontiers in Neuroinformatics*, 9, 9 2015. ISSN 1662-5196. doi: 10.3389/fninf.2015.00023. URL http://journal.frontiersin.org/Article/10.3389/fninf.2015.00023/abstract.

Siyuan Gao, Abigail S. Greene, R. Todd Constable, and Dustin Scheinost. Combining multiple connectomes improves predictive modeling of phenotypic measures. *NeuroImage*, 201:116038, 11 2019. ISSN 10959572. doi: 10.1016/j.neuroimage.2019.116038.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. Allennlp: A deep semantic natural language processing platform. 2017.

Matthew F. Glasser, Stamatios N. Sotiropoulos, J. Anthony Wilson, Timothy S. Coalson, Bruce Fischl, Jesper L. Andersson, Junqian Xu, Saad Jbabdi, Matthew Webster, Jonathan R. Polimeni, David C. Van Essen, and Mark Jenkinson. The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*, 80:105–124, 10 2013. ISSN 1053-8119. doi: 10.1016/J.NEUROIMAGE.2013.04.127. URL https://www.sciencedirect.com/science/article/pii/S1053811913005053?via%3Dihub.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. Thinking ahead: prediction in context as a keystone of language in humans and machines. *bioRxiv*, pages 2020–12, 2021.

Abigail S. Greene, Siyuan Gao, Dustin Scheinost, and R. Todd Constable. Task-induced brain state manipulation improves prediction of individual traits. *Nature Communications*, 9(1):1–13,

12 2018. ISSN 20411723. doi: 10.1038/s41467-018-04920-3. URL www.nature.com/naturecommunications.

Ludovica Griffanti, Gholamreza Salimi-Khorshidi, Christian F. Beckmann, Edward J. Auerbach, Gwenaëlle Douaud, Claire E. Sexton, Eniko Zsoldos, Klaus P. Ebmeier, Nicola Filippini, Clare E. Mackay, Steen Moeller, Junqian Xu, Essa Yacoub, Giuseppe Baselli, Kamil Ugurbil, Karla L. Miller, and Stephen M. Smith. ICA-based artefact removal and accelerated fMRI acquisition for improved resting state network imaging. *NeuroImage*, 95:232–247, 7 2014. ISSN 10959572. doi: 10.1016/j.neuroimage.2014.03.034. URL /pmc/articles/PMC4154346/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4154346/.

Liberty S Hamilton and Alexander G Huth. The revolution will not be controlled: natural stimuli in speech neuroscience. *Language, Cognition and Neuroscience*, 35(5):573–582, 2020.

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. ISSN 00368075. doi: 10.1126/science.1089506. URL http://science.sciencemag.org/.

Martin N Hebart, Brett B Bankson, Assaf Harel, Chris I Baker, and Radoslaw M Cichy. The representational dynamics of task and object processing in humans. *Elife*, 7:e32816, 2018.

Anne Hsu, Alexander Borst, and Frédéric E Theunissen. Quantifying variability in neural responses and its application for the validation of model predictions. *Network: Computation in Neural Systems*, 15(2):91–109, 2004.

Alexander G Huth, Shinji Nishimoto, An T Vu, and Jack L Gallant. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. 2012. doi: 10.1016/j.neuron.2012.10.014. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3556488/pdf/nihms418681.pdf.

Alexander G Huth, Wendy A de Heer, Thomas L Griffiths, Frédéric E Theunissen, Jack L Gallant, Wendy a De Heer, Thomas L Griffiths, and Jack L Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, 2016. doi: 10.1038/nature17637. Natural.

Shailee Jain and Alexander Huth. Incorporating context into language encoding models for fmri. In *Advances in neural information processing systems*, pages 6628–6637, 2018.

Mark Jenkinson, Peter Bannister, Michael Brady, and Stephen Smith. Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. *NeuroImage*, 17 (2):825–841, 10 2002. ISSN 10538119. doi: 10.1006/nimg.2002.1132.

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

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.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 11 2008. ISSN 16625137. doi: 10.3389/neuro.06.004.2008. URL http://journal.frontiersin.org/article/10.3389/neuro.06.004.2008/abstract.

Stephen W Kuffler. Discharge patterns and functional organization of mammalian retina. *Journal of neurophysiology*, 16(1):37–68, 1953.

Mark D Lescroart and Jack L Gallant. Human scene-selective areas represent 3d configurations of surfaces. *Neuron*, 101(1):178–192, 2019.

T.M. Mitchell, S.V. Shinkareva, A. Carlson, K.M. Chang, V.L. Malave, R.A. Mason, and M.A. Just. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880): 1191–1195, 2008.

Rosaleena Mohanty, William A. Sethares, Veena A. Nair, and Vivek Prabhakaran. Rethinking Measures of Functional Connectivity via Feature Extraction. *Scientific Reports*, 10(1):10, 12 2020. ISSN 20452322. doi: 10.1038/s41598-020-57915-w. URL https://www.nature.com/articles/s41598-020-57915-w.

Samuel A Nastase, Ariel Goldstein, and Uri Hasson. Keep it real: rethinking the primacy of experimental control in cognitive neuroscience. *NeuroImage*, 222:117254, 2020.

S. Nishimoto, A.T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J.L. Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 2011.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

Aniketh Janardhan Reddy and Leila Wehbe. Syntactic representations in the human brain: beyond effort-based metrics. *bioRxiv*, 2020.

Monica D. Rosenberg, Emily S. Finn, Dustin Scheinost, Xenophon Papademetris, Xilin Shen, R. Todd Constable, and Marvin M. Chun. A neuromarker of sustained attention from whole-brain functional connectivity. *Nature Neuroscience*, 19(1):165–171, 12 2015. ISSN 15461726. doi: 10.1038/nn.4179. URL https://www.nature.com/articles/nn.4179.

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, 4 2014. ISSN 10538119. doi: 10.1016/j.neuroimage.2013.11.046. URL /pmc/articles/PMC4019210/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4019210/.

Martin Schrimpf, Idan Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative reverse-engineering converges on a model for predictive processing. *BioRxiv*, 2020.

X. Shen, F. Tokoglu, X. Papademetris, and R. T. Constable. Groupwise whole-brain parcellation from resting-state fMRI data for network node identification. *NeuroImage*, 82:403–415, 11 2013. ISSN 10538119. doi: 10.1016/j.neuroimage.2013.05.081. URL /pmc/articles/PMC3759540/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3759540/.

Saurabh Sonkusare, Michael Breakspear, and Christine Guo. Naturalistic stimuli in neuroscience: critically acclaimed. *Trends in cognitive sciences*, 23(8):699–714, 2019.

Gustavo Sudre, Dean Pomerleau, Mark Palatucci, Leila Wehbe, Alona Fyshe, Riitta Salmelin, and Tom Mitchell. Tracking neural coding of perceptual and semantic features of concrete nouns HHS Public Access. *Neuroimage*, 62(1):451–463, 2012. doi: 10.1016/j.neuroimage.2012.04.048.

Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Advances in Neural Information Processing Systems*, pages 14928–14938, 2019.

Mariya Toneva, Tom M Mitchell, and Leila Wehbe. Combining computational controls with natural text reveals new aspects of meaning composition. *bioRxiv*, 2020.

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. ISSN 1053-8119. doi: 10.1016/J.NEUROIMAGE.2013.05.041. URL https://www.sciencedirect.com/science/article/pii/S1053811913005351?via%3Dihub.

Tyler J VanderWeele. Outcome-wide epidemiology. *Epidemiology (Cambridge, Mass.)*, 28(3):399, 2017.

Aria Y Wang, Leila Wehbe, and Michael Tarr. Neural taskonomy: Inferring the similarity of task-derived representations from brain activity. 2019.

Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. Simultaneously Uncovering the Patterns of Brain Regions Involved in Different Story Reading Subprocesses. *PLoS ONE*, 9(11):e112575, 11 2014a. ISSN 1932-6203. doi: 10.1371/journal.pone.0112575. URL https://dx.plos.org/10.1371/journal.pone.0112575.

Leila Wehbe, Ashish Vaswani, Kevin Knight, and Tom Mitchell. Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 233–243, 2014b.

Leila Wehbe, Idan Asher Blank, Cory Shain, Richard Futrell, Roger Levy, Titus von der Malsburg, Nathaniel Smith, Edward Gibson, and Evelina Fedorenko. Incremental language comprehension difficulty predicts activity in the language network but not the multiple demand network. *Cerebral Cortex*, 31(9):4006–4023, 2021.

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.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.