Improved Mapping of Information Processing in the Brain During Naturalistic Experiments

Anand Bollu

CMU-CS-21-124

Aug 2021

Computer Science Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Leila Wehbe, Chair
Michael J. Tarr

Submitted in partial fulfillment of the requirements for the degree of Master of Science.

Keywords: neuroscience, brain mapping, encoding models, language, fMRI

Abstract

Modern cognitive neuroscience has focused on mapping what information is processed both where and when in the brain as an initial step towards understanding *how* information is processed by the brain. Brain mapping is most commonly performed using the linear encoding model - a computational tool that can be used to align brain measurements with numerical stimulus representations that can be informative of the underlying neural processes. But as we move towards more naturalistic experimental settings and start to work with increasingly complex stimulus representations, it becomes important to test the boundaries of what we are able to learn using existing computational tools so we can adjust them accordingly. In this work, we examine two potential limitations of the linear encoding model. First, we look at a specific problem setting that shows that the existing framework does not always allow us to disentangle where specific stimulus-related information is processed in the brain. As an initial step towards addressing this issue, we propose a new framework that can be used to group together areas of the brain that are responsible for processing similar information. Second, we explore whether using a more expressive, nonlinear encoding model can allow us to better align the internal representational spaces of artificial neural networks, from which we can derive more complex and potentially more informative stimulus representations, with those of the brain. Overall, our proposed extensions to conventional encoding approaches can help neuroscientists unlock a richer and more complete space of brain mappings.

Acknowledgments

I would like to thank Leila Wehbe and Mariya Toneva for their invaluable support and guidance over the last two years. From taking the time to help me design experiments and interpret results to nudging me in the right direction when it came time to think about my future plans, they have gone above and beyond to help me reach my goals. I'd also like to thank previous and present members of our research group: Jennifer Williams (who is also a co-author on one of the works included in this thesis), Ruogu Lin, Aria Wang, Aniketh Reddy, Srinivas Ravishankar, Maggie Henderson and Nidhi Jain. Thank you all for bringing exciting ideas to our weekly group meetings and providing many insightful suggestions that have helped refine the works included in this thesis. Lastly, I would like to thank my sister, my parents and my friends for being endless sources of happiness and support during these past few years.

Contents

List of Figures

Chapter 1

Introduction

Encoding models have emerged as the computational tool of choice for researchers interesting in mapping where specific information is processed in the brain. Recent work has shown that the linear encoding model, which predicts brain activity as a linear combination of a numerical stimulus representation, can be used to reveal important properties of information processing in the brain, such as that the representation of concepts is distributed across the cortex but consistent across people [33, 34, 39, 43, 46, 65]. In these cases, researchers used interpretable representations of the stimuli to link specific information with the location where this information is processed in the brain (i.e. a concrete noun stimulus was represented as its co-occurrence with a set of verbs in a large text corpora [43]).

But with the advent of machine learning in recent years, researchers have started to use linear encoding models in conjunction with more complex representations derived from artificial neural networks (ANNs) as a way to understand where the brain may be representing high level semantic information about the stimulus. Since concerted efforts have succeeded at building ANNs that can match human-level performance on specific tasks like language comprehension and image understanding, these networks and the representations within seem to capture useful semantic information about their inputs. For example, language models have been shown to contain information about part of speech [42] and semantic roles [59]. While these complex stimuli representations are able to predict brain measurements to an unprecedented extent [6, 7, 25, 37, 54, 60, 64, 67, 72], their complexity also makes it more difficult to make scientific inferences about what specific information is processed where in the brain. This inference becomes even more difficult in experimental settings where participants observe naturalistic stimuli (e.g. watching movies, reading books, listening to stories), which are becoming increasingly more popular in neuroscience [28, 45, 57]. Although this naturalistic setting enables studying processing that is more easily generalizable to the everyday world, it can further complicate brain mapping because there is less control over what information the brain is actually processing about the stimulus.

In this thesis, we delve into two potential shortcomings of the current encoding model framework that may be holding it back from aligning information from complex representations of naturalistic stimuli with brain activity. First, we investigate a specific naturalistic problem set-

ting that highlights why existing computational approaches may fail to disentangle where specific stimulus-related information is processed in the brain. As an initial step towards overcoming these challenges, we put forth a promising new tool that can help us identify groups of regions in the brain where similar stimulus-related information is processed. Second, we show that adding complexity beyond the commonly used linear encoding model might allow us to deal with the added complexity in modern stimulus representations.

Understanding what information is processed both where and when in the brain helps us get closer to answering *how* it is processed by the brain. In this thesis, we mainly focus on the what and where questions. Since we rely mostly on fMRI data, which has high spatial resolution and low temporal resolution, for all experiments that are part of this work, we do not address the when question here. Also, the works included primarily focus on language processing in the brain but the ideas we introduce are general and can be applied to a wide variety of stimulus representations.

1.1 Overview of Methods

fMRI data: Hidden Figures. The experiments we perform in Sections 2 and 3 use fMRI recordings of 6 healthy participants presented with the movie *Hidden Figures* in English, made available by the Courtois Neuromod group (data release `cneuromod-2020`). The movie was split up into roughly 10 minute segments that were presented to participants in separate runs. A total of approximately 120 minutes of data were recorded for each participant, at a repetition time (TR) of 1.49 seconds. Apart from brain data, this dataset also provides word-level timestamp estimates for every movie segment. Since words are presented at an uneven rate in this setting, these timestamps play a crucial role in allowing us to group words based on the TR interval they appear in. This data is available upon request at `https://docs.cneuromod.ca/en/latest/ACCESS.html#downloading-the-dataset`

fMRI data: HCP short movies. We use publicly available data from the Human Connectome Project (HCP) 7T dataset, with healthy participants between 22-36 years old [63]. HCP fMRI data comes minimally pre-processed as FIX-Denoised data [24, 27, 53]. We focus our analysis on only 90 participants for now, and are not using the remaining due to another project. Each participant watched naturalistic audio-visual video clips in English during 4 scans. Each scan was just over 15 minutes long, 60 minutes and 55 seconds of data were recorded. The fMRI sampling rate (TR) was 1 second.

fMRI data: Harry Potter and the Sorcerer's Stone. We use fMRI recordings of 8 healthy participants reading chapter 9 of Harry Potter and the Sorcerer's Stone [52]. This data was made available by Wehbe et. al [66]. Each word was presented for 0.5 seconds and approximately 45 minutes of data were recorded for each participant, at a repetition time (TR) of 2 seconds.

Training encoding models. Given TR-level stimulus representations $[z_0, z_1, \ldots z_T]$ (where T is the total number of TR intervals), we train encoding models to predict voxel activities at the j-th

TR interval using a concatenated vector of stimulus representations corresponding to previous intervals [z_{j-1}, z_{j-2}, z_{j-3}, z_{j-4}]. This lets us account for the lag in the hemodynamic response that is characteristic of fMRI recordings. Similar to previous work [34, 46, 58, 60, 65, 67], we use ridge-regression and k-fold cross-validation (we specify k separately for each study) during training. We use nested 10-fold cross-validation to choose regularization parameters independently for each voxel.

Chapter 2

Case study: encoding models in movie watching

2.1 Introduction

In this chapter, we look at an exemplar problem setting that helps shed light on some of the challenges of working with complex representations of naturalistic stimuli. We are specifically interested in disambiguating where in the brain high-level semantic information is processed from where potential confounders like speaker identity and word rate are represented when watching a movie. Our approach uses linear encoding models to examine how well a variable or representation that captures specific stimulus-related information can significantly predict brain activity in different regions. Here, we use internal representations from a bidirectional language model called ELMo [49] as proxies for some of the high-level semantic information that is also represented in the brain. Like its contemporaries, ELMo can be optimized to perform well on a variety of downstream linguistic tasks so these representations seem to capture something generic about the given language input.

2.2 Methods

fMRI data. We use fMRI recordings of 6 healthy participants presented with the movie *Hidden Figures* in English, made available by the Courtois Neuromod group. For more details, refer to *fMRI data: Courtois Neuromod* in Section 1.1.

Stimulus representations. We extract three types of representations as proxies for semantic, speaker-specific and word rate information from the stimulus. For high-level semantic information, we use word-level representations from the first hidden layer of a pretrained version of ELMo [23] (a bidirectional language model), obtained by passing the previous 25 words to the network. Prior work has shown that ELMo representations collected at this context length and layer depth can be significantly predictive of fMRI recordings of participants comprehending language [60, 61]. We also extract our own word-level speaker labels by aligning the word-by-word speech transcriptions provided by the dataset with subtitles that map between speakers and

their dialogues. The collected speaker labels span 10 of the movie's characters that have the most dialogues. The remaining characters are grouped under the `OTHER` label. From a total of approx. 11000 words, 2 speakers are mapped to approx. 2000 words each, another 2 to approx. 1000 words each and the remaining 6 to approx. 200-700 words each. The rest of the words (approx. 2500) are annotated with the `OTHER` label. Using these labels, we construct word-level one-hot speaker representations. We use timestamp estimates from the dataset to transform the word-level ELMo and speaker representations to TR-level representations by averaging those that appear together in the same TR interval. Finally, we also generate our own TR-level word rate labels that indicate the number of words presented in each interval.

Encoding models. First, we estimate encoding models which predict brain activity from each stimulus representation directly. For details about model training, refer to Section 1.1. All encoding models here are trained using 12-fold cross-validation. What we find is that ELMo, word rate and speaker identity predict similar areas of the brain similarly well so in an effort to disambiguate the unique underlying neural processes related to each of them, we further estimate encoding models which predict brain activity after specific stimulus-related information has been regressed out from an existing representation. To regress out information Y from a stimulus representation X, we generate predictions $\hat{X}$ from Y and collect residual representations, $X_{\_Y} = X - \hat{X}$. This should allow us to compare how encoding performance is affected when information Y is no longer present in the encoding model's input space. More information about model evaluation and these experiments can be found in section 2.3.

2.3 Results

ELMo, word rate and speaker identity predict the brain similarly well. We train separate encoding models to predict brain activity from each stimulus representation. We then evaluate encoding performance by computing voxel-level Pearson correlations [17] between the models' predictions and the true activities from held-out data. Using voxel-level permutation tests (significance level 0.05, FDR controlled for multiple comparisons [2]), we determine the statistical significance of each model's performance independently for each subject. Lastly, we use pycortex [21] to visualize the number of subjects where each voxel is significantly predicted, which we show in Fig. 2.1. Here, white voxels are predicted significantly from the input stimulus representation for most subjects. On the other hand, darker (blue) voxels could not be predicted significantly well from the same representation by our encoding model. What we find is that all three stimulus representations are able to consistently predict similar clusters of voxels, mainly from the temporal and parietal lobes, similarly well.

Effect of regressing out word rate. Next, we investigate whether word rate might be a confound in trying to understand how semantic information captured by ELMo and speaker-specific information are represented in the brain. We first verify that ELMo representations and speaker identity can significantly predict word rate (permutation test, significance level 0.05) to confirm that both of them indeed interact with word rate. After confirming this, we regress out word rate information from each of the two representations and estimate separate encoding models

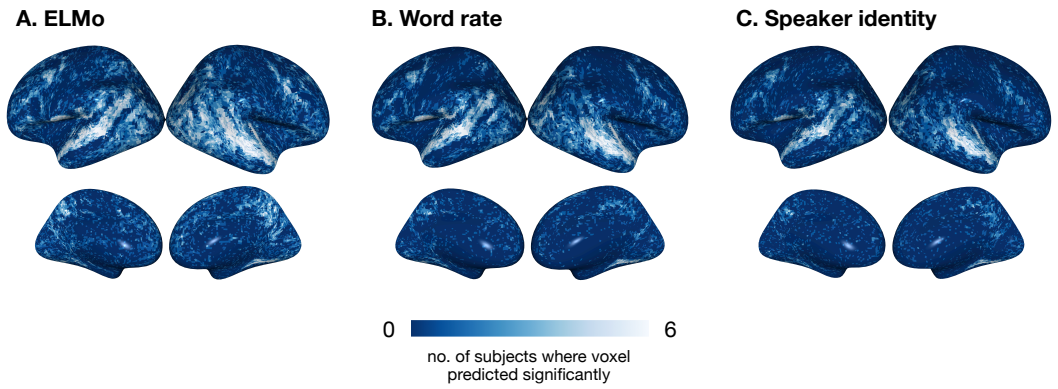

Figure 2.1: Plotting number of subjects where each voxel is predicted significantly by ELMo, word rate and speaker identity.

to predict brain activity using the resulting residual representations. We compute the difference between the voxel-level Pearson correlations collected before and after word rate is regressed out and perform permutation tests (significance level 0.05, FDR controlled for multiple comparisons) to determine whether this difference is significant for each voxel. Again, we use pycortex to plot the number of subjects where a significant difference is found, but restrict our visualization to only those voxels that were predicted significantly by the original representation (ELMo or speaker identity) for 3 or more of the 6 participants. The mass of white voxels we see across the temporal lobe in Fig. 2.2 shows that there is a significant difference in encoding performance in this region across a majority of participants when word rate information is regressed out from both ELMo and speaker identity. This seems to suggest that the alignment we found earlier in these clusters may largely be attributed to the processing of low-level information related to word rate here.

Effect of regressing out speaker identity. We use a similar approach to also investigate whether speaker identity might be a confound in trying to understand how other information captured by ELMo is represented in the brain. Again, we verify that ELMo representations can significantly predict speaker identity (permutation test, significance level 0.05, FDR controlled for multiple comparisons). We find that 4 out of 5 speakers with over 1000 words over the whole movie and 4 out of the remaining 6 speaker labels can be predicted significantly from ELMo, which confirms that they both interact with each other. We then perform voxel-level permutation tests (significance level 0.05, FDR controlled for multiple comparisons) to determine whether regressing out speaker information leads to a significant difference in encoding performance. Fig. 2.3A shows a significant difference in encoding performance, again in the temporal lobe, across a majority of participants when speaker information is regressed out from both ELMo. However, our results from the previous section already indicated that word rate might be a potential confounder in understanding how the brain processes both ELMo- and speaker-related information. So we also regress out speaker information from the residual representation that is obtained by first regressing out word rate from ELMo. The dark blue clusters we observe in Fig. 2.3B suggests that there are few significant differences in encoding performance after word rate is accounted for.

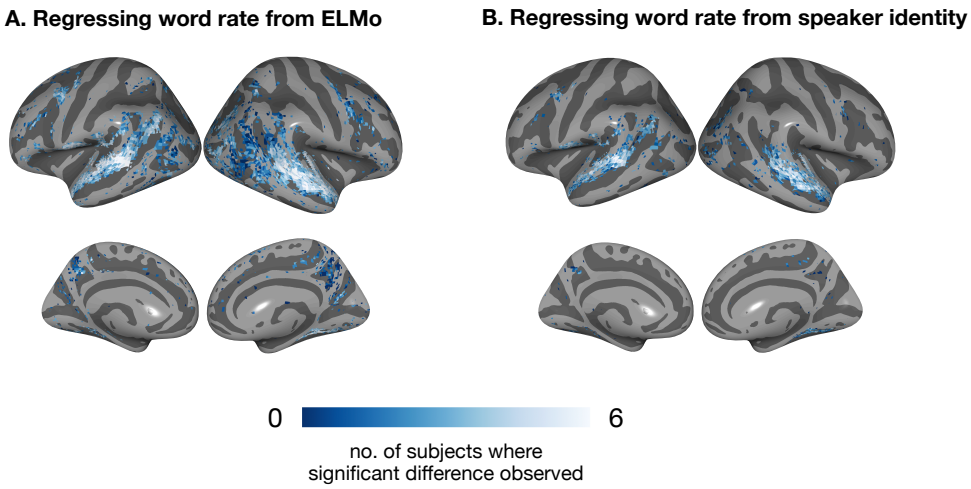

Figure 2.2: Number of subjects where we find a significant difference in encoding performance after word rate information is regressed out from ELMo and speaker identity.

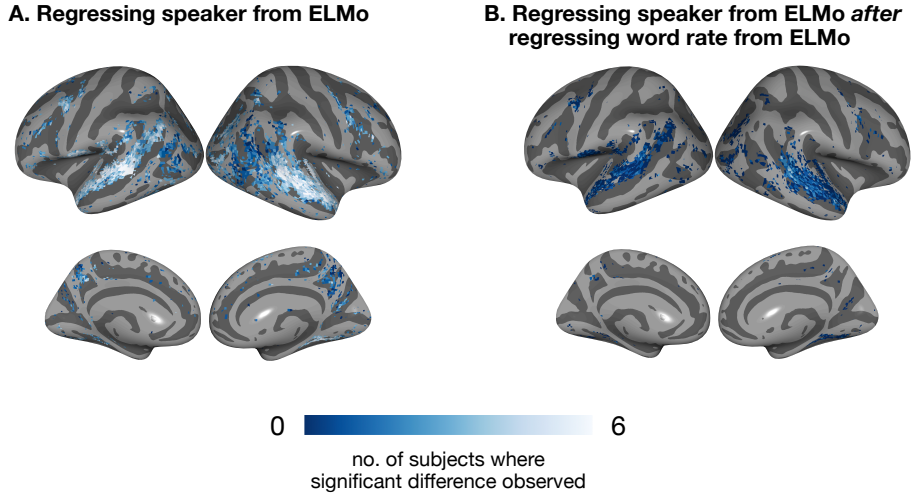

Figure 2.3: Number of subjects where we find a significant difference in encoding performance after (A) speaker information is regressed out from ELMo and (B) speaker information is regressed out from ELMo after first regressing out word rate information.

2.4 Discussion

We examine what linear encoding models can tell us about where the brain processes specific information when presented with a complex naturalistic stimulus - here, the movie *Hidden Figures*. We observe that word rate, speaker identity and ELMo representations can predict brain activity in similar regions similarly well. Our findings suggest that word rate acts as a confound in trying to understand where the brain processes ELMo- and speaker-specific information. We specifically find that parts of the temporal lobe seem to respond to low-level word rate-related information by regressing out word rate from ELMo and from speaker identity. However, a similar approach does not allow us to separate where the brain processes speaker information from where it processes all other information captured within ELMo representations. This suggests that regressing out a potential confound from a stimulus representation of interest may not always allow us to identify where specific information is processed. To harness the wealth of stimulus-related information that ANN representations seem to capture about naturalistic stimuli, we need a more reliable way to map out what information is processed where in the brain. In the next chapter, we start to fill this gap by introducing two new metrics and an analysis framework that, when used together, can allow us to infer a more specific relationship among brain regions with respect to the stimulus.

Chapter 3

Improved scientific inference for encoding models of complex stimuli

This is joint work with Mariya Toneva, Jennifer Williams, Christoph Dann and Leila Wehbe. My main contribution was to lead the simulations component of this project - I used simulated data with known ground truth relationships between brain measurements, an observed stimulus and a stimulus representation to highlight the limitations of the current encoding model framework that we rely on for inferring these relationships (section 3.4 and appendix section A.4).

3.1 Introduction

As we move towards more complex stimuli representations and more complex experimental settings, we need to reexamine what inferences we're able to make from our existing computational tools and adjust our toolbox accordingly. Here, we present a detailed analysis of what scientific inferences we're able to make using the linear encoding model and propose two new tools that, when used together, can allow us to recognize areas of the brain that are involved in processing similar information.

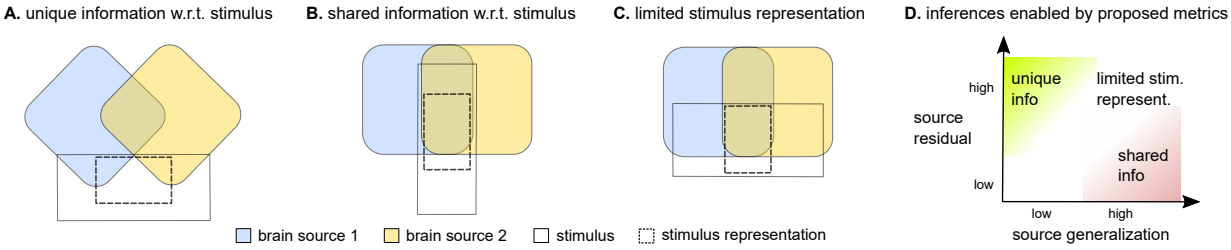

Figure 3.1: Venn diagrams representing different cases for the underlying relationships between two brain measurements, the presented stimulus, and the stimulus representation (A-C), and how our proposed metrics enable us to infer these relationships (D). In contrast, in each case an encoding model will predict a similar proportion of variance in both brain measurement sources, making it difficult to disambiguate the three cases.

Motivating example A simple conceptual example illustrated in Fig. 3.1 (A-C) motivates the need to reexamine the interpretation of encoding model performance when using complex stimuli representations. We present three different cases for the underlying relationships between two brain measurements shown in yellow and in blue, the stimulus corresponding to these measurements, and the stimulus representation that is used to train an encoding model. In the first case (Fig. 3.1A) the two brain measurement sources process unique stimulus information. For instance, the blue source may capture the part of speech of a stimulus word and the yellow source may capture some of its semantic properties (e.g. manipulability and size). The stimulus representation also captures these multiple types of information (e.g. the stimulus representation may be derived from a language model, such as ELMo [49]), so an encoding model using this stimulus representation would predict a proportion of the variance in both sources. In this case, the encoding model performance would mislead us to think that the two brain sources process the same information about the word stimulus.

In the second and third cases (Fig. 3.1B-C) the two brain sources share information with respect to the stimulus, as indicated by the common overlap between the yellow and blue brain sources and the large stimulus rectangle. However, in Fig. 3.1B each brain source captures little unique stimulus information, while in Fig. 3.1C each source captures a lot of unique information. In both of these cases, using the same stimulus representation as input to an encoding model would predict both brain sources to a similar degree, which limits our ability to disentangle the two possible cases. This limitation is due to the limited stimulus representation in case C, which only captures aspects of the shared information. This is an issue because we can only make the claim that a set of regions is processing the same stimulus information in the case that we are able to disentangle Case B from the rest. To address these limitations, we propose two new metrics that, when used together, can disambiguate each of the three cases, as discussed more in depth in Section 3.3.

Real-world example of inference problem. One case study of this inference problem from the neuroscience literature is a set of findings using encoding models during naturalistic language comprehension. Several researchers have found that the activity in a wide set of bilateral regions in the temporal and prefrontal cortices deemed as the language network [15] can be significantly predicted as a function of various features of the presented language stimuli [5, 50, 65], without distinguishing between these regions in terms of the information map. [10, 34] do show that the set regions predicted by word meaning are tuned to different aspects of meaning, however, the set of regions in the language network are shown to be tuned to the same aspects of meaning. A big outstanding question from these works is whether these language regions are all indeed processing the same information about the stimulus, or whether the tools we are using (e.g. stimulus representations, encoding models) and the way we are using them (e.g. by interpreting the encoding model performance) is preventing us from differentiating between them.

In this work, we first highlight limitations of existing computational approaches for this scientific inference using simulated data with known ground truth relationships. We then propose two new metrics and an analysis framework that, when used together, can allow us to make stronger scientific inferences that can disambiguate all three cases. We term these metrics *source gen-*

eralization and *source residuals*. We next use our proposed tools to analyze two fMRI datasets obtained using naturalistic stimuli. Our main contributions are as follows. We conceptually breakdown the possible underlying relationships between two brain sources and the presented stimuli when both sources are predicted by the same feature space and present two metrics that can distinguish them. We present simulations that show limitations of commonly used methods for disambiguating these different relationships. Finally, we showcase the use of these metrics in two fMRI datasets with naturalistic stimuli showing when we can disambiguate and when the feature set is the limitation. Our results generalize across these two datasets that capture different populations and are acquired by different labs in different countries with very different experimental setups and scanning parameters.

3.2 Related Work

Encoding models are becoming popular as datasets with complex stimuli become more common and stimulus representations from neural networks are employed as a tool to study the brain. Several works have investigated the expressivity of encoding models. Wu et al. [71] frames brain mapping as system identification, where the neuroscientist is attempting to discover the features that different parts of the system are sensitive to. Wu et al. [71] also describes a methodology for building encoding models and measuring the performance ceiling.

Some work has described the utility of encoding models as opposed to decoding models, in making precise inferences about representations [44, 70] or have proposed ways to make decoding models more interpretable [30]. Some work has used some version of variance partitioning to differentiate between *feature spaces* [5, 9, 41, 50, 61]. However, here we focus on encoding models with the *same feature space* at *different sources* (e.g. different voxels, different neurons, different regions, or different sensors). Some work has relied on the correlation between voxels to enforce priors on the learned models [48, 68]. Other than pooling information across voxels to regularize encoding models, most work in relating the information in different voxels has been in the functional connectivity literature [62]. In that large body of work, there is not usually a relationship drawn between the stimulus and the brain activity. Instead, the activity between different sources is correlated, often during rest, and the correlation used as a metric for functional connectivity. Some work has used inter-subject correlation to identify which regions are related to a stimulus [29, 56]. However, we focus on understanding the relationship between regions and the stimulus.

Because of the high temporal resolution of certain recording tools like magnetoencephalography (MEG), it is possible to compare the representations across time points. A powerful method called temporal generalization has been proposed by King and Dehaene [40]. It allows researchers to see if the representation at a given point in time is similar to another point in time, and have been subsequently used in multiple works [4, 19, 20, 31]. The generalization idea was recently adapted by [61] into spatial generalization, in which the representation at different voxels are compared. This is done by using the predictions at one voxel to predict the activity in another voxel. These methods can be seen as specific instantiations of the *source generalization*

metric that we describe. We further propose a new way to normalize this source generalization metric which improves its interpretability.

Another approach is to compare encoding model weights, which has typically been done after reducing the dimensionality across the brain and plotting the resulting low dimensional projections on the brain [10, 33, 34]. In Çukur et al. [8], different encoding models are estimated for the same participant under different attention conditions, and the tuning change due to attention is estimated from the weights of the different models.

3.3 Definitions

In this section, we define our two proposed metrics and other existing metrics that we commonly reference throughout this work. To ground these definitions, we relate each metric to an underlying setting for how a brain measurement is assumed to be generated. We introduce three such settings that are increasingly more specific about the dependence between one brain source and other sources in the same participant and other participants. Each setting introduces an additional set of assumptions, which enable us to relate the metrics to the unique and shared information in each brain source.

3.3.1 Single brain source, single participant

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

$$Y = g(X) + \epsilon \tag{3.1}$$

where $Y \in$ corresponds to the observation at a single brain source (e.g. fMRI voxel, EEG/MEG sensor-timepoint, electrode), $X \in^d$ to 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 $\epsilon \in$ to a noise term that is independent from the stimulus representations.

Encoding model. The first setting is the one that commonly underlies encoding models. In this setting, an encoding model estimates the function $g(\cdot)$ for a specific stimulus representation X. Most commonly, this function is parameterized as a linear function (i.e. $g(X) = \langle X, \theta \rangle$, where $\theta \in R^d$), and is estimated using a set of training observations for each brain source in a cross-validated fashion. For a set of predictions Y on heldout data, the encoding model performance is defined as follows:

$$\texttt{encoding model performance}(Y, Y) = (Y, Y)\,,$$

where $(\cdot)$ signifies Pearson correlation.

3.3.2 Multiple brain sources, single participant

In the second setting, we explicitly allow for a noise term reflecting shared information between two brain sources Y_1 and Y_2 that is not captured by the stimulus representation. In our modified setting,

$$Y_i = \underbrace{g_i(X)}_{\text{signal}} + \underbrace{\epsilon_i}_{\text{individual noise}} + \underbrace{\epsilon_{12}}_{\text{shared noise}} \tag{3.2}$$

for $i \in \{1, 2\}$ where $g_i(X) = \langle X, \theta_i \rangle$ are linear functions of stimulus representations $X \in R^d$ with parameters $\theta_i \in R^d$ and $\epsilon_1, \epsilon_2, \epsilon_{12}$ are noise terms. All noise terms are independent of each other and X and have zero mean and variances σ_1^2, σ_2^2 and σ_{12}^2, respectively.

Functional Connectivity. Functional connectivity measures the correlation of two brain sources Y_1 and Y_2 through time. In the setting above, this evaluates to

$$(Y_1, Y_2) = \frac{\text{cov}(Y_1, Y_2)}{\sqrt{(Y_1)(Y_2)}} = \frac{\text{cov}(g_1(X), g_2(X)) + \sigma_{12}^2}{\sqrt{(Y_1)(Y_2)}} \,,$$

where $(Y_i) = (g_i(X)) + \sigma_i^2 + \sigma_{12}^2$ is the variance of each observation. Note that (Y_1, Y_2) does not distinguish between stimulus-related and stimulus-independent components of Y_i. That is, a correlation of Y_1 and Y_2 can be high if either $\text{cov}(g_1(X), g_2(X))$ or σ_{12}^2 is large. We support this observation using simulated brain source data in Section 3.4.

First proposed metric: source generalization. Disambiguating the three possible cases presented in Fig. 3.1A-C is key to conclusively inferring whether two sources process the same information about a stimulus. As outlined in Section 1, encoding performance is not able to disambiguate any of these three cases. To address some of these limitations, we propose our first metric *source generalization*. Intuitively, source generalization captures the amount of information shared by two sources and the stimulus representation used in an encoding model. Concretely, we define source generalization as:

$$\texttt{source generalization}(Y_1, Y_2) = (Y_1, Y_2) \,, \tag{3.3}$$

where Y_1 is a prediction on heldout data obtained by an encoding model as defined in Section 3.3.1.

Source generalization helps to disentangle the case shown in Fig. 3.1A from the cases in B and C. The source generalization will be low in case A because there is no shared information between the two brain sources and the stimulus representation, so training an encoding model on brain source 1 will not generalize to brain source 2, and vice versa. In contrast, the source generalization will be high in cases B and C because there is shared information between the two brain sources and the stimulus representation. However, source generalization is not able to disambiguate case B from case C.

3.3.3 Multiple brain sources, multiple participants

To disambiguate case B from case C, we consider our third setting which allows for multiple participants $P \in \{A, B\}$ and for each quantity in Equation (3.2) to differ in each subject P:

$$Y_{i,P} = \underbrace{g_{i,P}(X)}_{\text{signal of stimulus representation}} + \underbrace{h_i(Z)}_{\text{signal of } Z} + \underbrace{\epsilon_{i,P}}_{\text{individual noise}} + \underbrace{\epsilon_{12,P}}_{\text{joint noise}}. \tag{3.4}$$

We additionally allow for the presence of an additive component $h_i(Z)$ which makes explicit a dependence on features Z that are part of the stimulus, but are not captured by the stimulus representation X. This enables us to express the likely common occurrence that a specific stimulus representation does not perfectly reflect all brian-relevant stimulus information.

Second proposed metric: source residuals. While source generalization is able to disentangle case A in Fig.3.1 from cases B and C, it is unable to disentangle case B from case C. To address this limitation, we introduce the second metric–*source residuals*. Here, we build on the intuition behind intersubject correlation to estimate how much of the information that is shared between two brain sources and the stimulus is *not* shared between the two brain sources. More concretely,

$$\texttt{source residual}(Y_1, Y_2) = (R_{1-2,A}, R_{1-2,B}), \tag{3.5}$$

where $R_{1-2,P} = Y_{1,P} - (Y_{1,P}, Y_{2,P})Y_{2,P}$ is the residual of regressing $Y_{2,P}$ from $Y_{1,P}$. Source residuals disentangle case B from case A and C. Thus, we argue that using both proposed metrics together can help disentangle each case from the others.

3.4 Simulations

Using simulated data, we look at what encoding model performance, functional connectivity, source generalization and source residuals can tell us about the underlying relationships between two sources.

Simulating stimulus information. We generate two components that together make up all available stimulus information: $X \in R^d$, which is the stimulus representation, and $Z \in R^d$, which contains the remaining stimulus information not in X. A key simplifying assumption we make when generating X and Z data is that both components can be decomposed into four disjoint independent subsets of stimulus information: unique information captured by the individual brain sources (X_1, X_2, Z_1, Z_2), joint information captured by both brain sources (X_{12}, Z_{12}) and information not captured by either brain source (X_3, Z_3). Each X_i, Z_i, of length $\frac{d}{4}$, is independently sampled from a multivariate normal with mean 0 and a symmetric toeplitz covariance matrix with diagonal elements equal to 1, and X and Z are constructed by concatenating their four corresponding sub-components.

Simulating brain source data. We simulate observations at two brain sources from two distinct participants using the following data generation model (motivated by Eq. 3.4):

$$Y_{i,P} = \alpha \times \underbrace{g_{12,P}(X)}_{\text{joint signal}} + (1-\alpha) \times \underbrace{g_{i,P}(X)}_{\text{unique signal}} + \alpha \times \underbrace{N_{i,P}}_{\text{unique noise}} + (1-\alpha) \times \underbrace{N_{12,P}}_{\text{joint noise}} \quad (3.6)$$

$$\text{where } N_{i,P} = \delta \times h_{i,P}(Z) + (1-\delta) \times \epsilon_{i,P} \text{ and } N_{12,P} = \delta \times h_{12,P}(Z) + (1-\delta) \times \epsilon_{12,P}.$$

Here, each $g_{i,P}(X) = \langle \theta_{i,P}, X_i \rangle$ is a linear function of the stimulus representation that only looks at the corresponding X_i in X. In order to generate the necessary participant-specific parameters $\theta_{i,P} \in R^{\frac{d}{4}}$, we first generate $\theta_i \in R^{\frac{d}{4}}$ by independently sampling each of its components from a uniform distribution over $[0, 1)$. Each $\theta_{i,P}$ is then sampled from $\mathcal{N}(\theta_i, 0.25\mathbf{I})$ to allow for variation between participants. The same approach is used to generate each $h_{i,P}(Z) = \langle \phi_{i,P}, Z_i \rangle$ term. $\epsilon_1, \epsilon_2, \epsilon_{12} \in R$ are terms that represent the information captured that is not related to the stimulus. Each ϵ_i is independently sampled from a standard normal distribution. See appendix for more details.

We introduce two adjustable parameters to simulate a wide range of scenarios. $\alpha \in [0, 1]$ controls how much of the stimulus representation related information captured by both brain sources is shared between them and how much of it is unique to each one. $\delta \in [0, 1]$ controls how much of the information captured by both brain sources that is unrelated to the stimulus representation is still related to the stimulus itself. It is important to note that the simulation results we present and analyze in this section are representative of an over-constrained setting. Their main purpose is to provide intuition for how the metrics discussed in Section 3.2 vary under different controlled scenarios.

Case A vs. Cases B & C. A key property that separates Case A from Cases B and C from Fig. 3.1 is the amount of unique stimulus information captured by each source that is also captured in the stimulus representation. We control this setting by keeping δ constant and varying α in Eq. 3.6. We use $\delta = 1.0$ here, but similar trends can be observed for any $\delta \in [0, 1]$. A low α simulates Case A as each brain source mostly captures unique stimulus information also present in the stimulus representation. A high α brings us closer to Cases B and C as this information is mostly shared between both brain sources for a given subject. We show the results in fig. 3.2 A. All metrics are collected and averaged across 1000 repetitions at each α. The encoding model performance, functional connectivity and source residuals remain relatively unchanged across the board. On the other hand, source generalization increases as we increase α. This suggests that looking at source generalization can allow us to recognize whether our data most resembles Case A or one of Cases B & C.

Case B vs. Case C. Cases B and C from Fig. 3.1 are similar in that the two sources share information with respect to the stimulus representation in both cases. However, they differ in the amount of unique information that each source captures about the stimulus. By setting $\alpha = 1.0$ and varying δ in Eq. 3.6, we can vary the amount of unique stimulus information that each source captures while preserving the amount of information that both sources share with respect to the stimulus representation. We show the results in fig. 3.2 B. At each value of δ, all met-

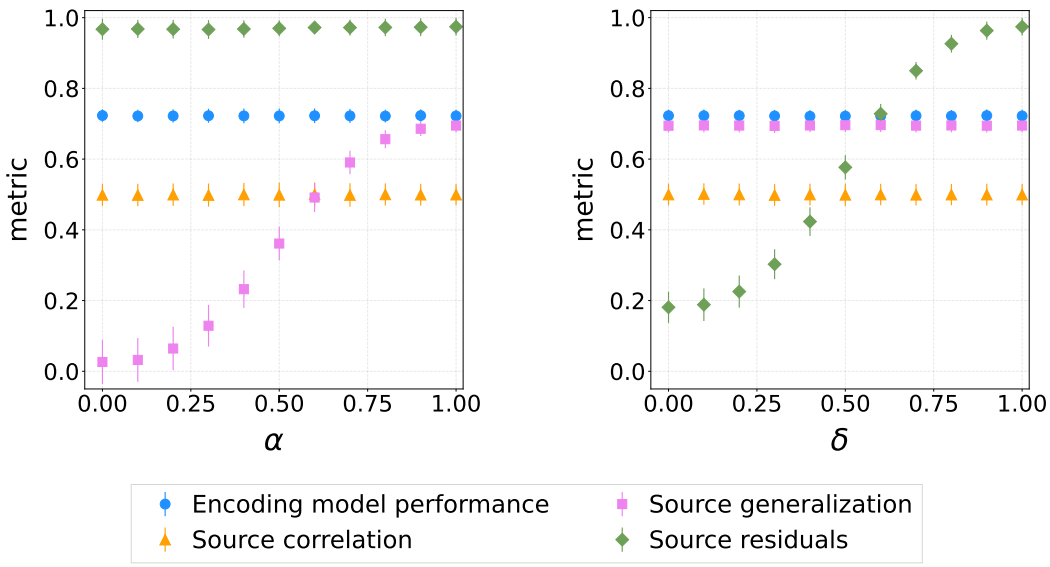

Figure 3.2: Plotting how each metric varies under simulations that separate (A) Case A from Cases B & C and (B) Case B from Case C.

rics are again collected and averaged across 1000 repetitions. The encoding model performance, functional connectivity and source generalization do not vary with respect to the δ used in our simulations. However, we observe that the source residuals increase with δ. These results suggest that once we identify that we are in either Case B or Case C, looking at the source residuals can help us recognize which of these cases our data most resembles.

3.5 Empirical Results on Two Naturalistic fMRI Datasets

We compute the quantities of interest for two fMRI datasets obtained when participants viewed naturalistic stimuli. We specifically chose these two datasets because they present a trade-off between the number of participants and the number of data recorded for each participant, which enables us to observe how the quantities of interest vary in real datasets with different numbers of samples.

fMRI data The first dataset we use contains fMRI recordings of healthy participants presented with naturalistic audio-visual clips from the Human Connectome Project (HCP) 7T dataset. The second fMRI dataset we use contains recordings of 6 healthy participants viewing the movie *Hidden Figures* in English, provided by the Courtois Neuromod group. For more details, refer to *fMRI data: HCP short movies* and *fMRI data: Courtois Neuromod* in Section 1.1.

Other data processing details. For each participant we downsample the fMRI data by averaging the voxel activities within the 268 functionally defined regions of interest (ROIs) from the Shen atlas per time frame [16, 55], similarly to previous work [12, 22, 26, 51]. For each participant this results in a dataset of dimensions - number of TRs by 268 ROIs. These ROIs

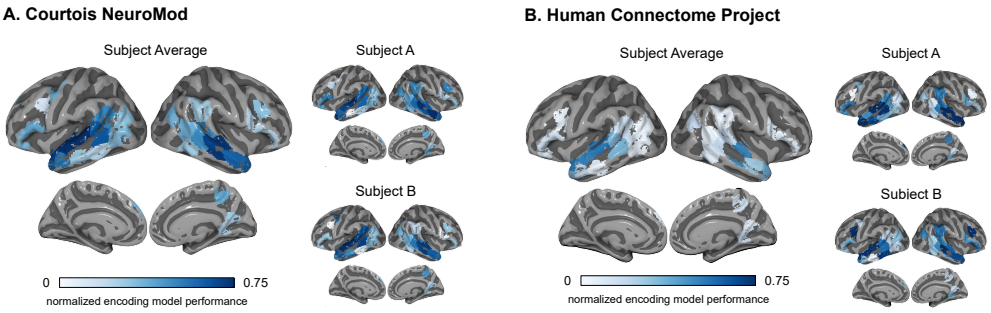

Figure 3.3: Encoding performance at 33 significantly predicted ROIs (corrected at level 0.05).

are entirely independent of our data as the Shen atlas was previously constructed from a separate group of healthy participants. The approach we propose apply to any brain regions. Because we are interested in studying naturalistic language comprehension, we chose to identify Shen atlas ROIs involved with processing language-relevant information. Regions of the brain involved with processing language-relevant information have previously been identified by [3, 15] and are also entirely independent of our data. We consider a Shen atlas ROI to be a language ROI if ¿ 15% of its voxels are within a region that processes language-relevant information. This procedure results in 55 Shen atlas ROIs that are language ROIs.

Stimulus representation. The approach we propose in this paper 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 use stimulus representations that capture the linguistic meaning of the stimuli. We follow previous work in neurolinguistics and obtain representations of the words observed by participants by feeding transcripts word-by-word into a pre-trained natural language processing model. We specifically choose ELMo [49], a bidirectional language model that incorporates multiple LSTM layers, for this purpose. Word representations obtained from the first hidden layer of ELMo, and contextualized with the previous 25 words, have been previously shown to significantly predict fMRI recordings of participants comprehending language [60, 61]. We focus our analyses on representations similarly collected from the first hidden layer of ELMo when provided with chunks of 25 consecutive words, using the pretrained ELMo provided by [23].

Encoding model performance. We first investigate what we can learn about how language regions process audio-visual stimuli by interpreting encoding model performance. We estimate encoding models which predict the brain activity associated with matching video clips from an ELMo embedding of the speech in the video clip. One encoding model is estimated independently for each participant's ROI. We provide details about model training in section 1.1. We use 4-fold cross-validation for HCP and 12-fold cross-validation for Courtois NeuroMod, which reflects the number of segments in which each dataset was originally collected. We evaluate the encoding model performance on heldout data for each fold.

In Figure 3.3, we present encoding model performances for the 33 language ROIs that were

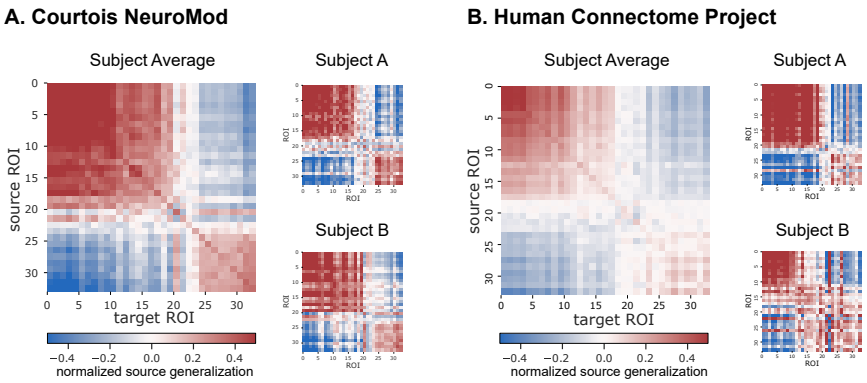

Figure 3.4: Source Generalization. ROI pairs with high norm. source generalization (red) process information captured by the stimulus representations in a similar way. Pairs with high norm. source generalization are consistent at the group and individual level in both datasets.

predicted significantly across participants in both fMRI datasets (one-sample t-test, FDR corrected for multiple comparisons across ROI at alpha level 0.05 [2]). The encoding model performances are normalized by the Intersubject Correlation (ISC), an estimate of the "noise ceiling" or the amount of variance in the ROI that is consistently related to the stimulus and therefore explainable (see Section 3.3.3 for a definition). We present the average normalized encoding performance across all participants in the datasets, as well as for two representative individual participants (see Appendix for additional individual-level performances). We observe that a set of bilateral language ROIs can be significantly predicted by the ELMo embedding for both datasets, at a group and individual participant level. These results replicate previous findings that the language ROI are well predicted by representations from ELMo [60, 61]. We also observe similarly to prior work that these regions are predicted significantly by the encoding model, thereby making it difficult to distinguish between them [5, 34, 50].

Functional connectivity. Next we investigate whether functional connectivity can disambiguate the regions that were found to be significantly predicted by ELMo representations. We present the pairwise functional connectivity for the 33 language ROIs in both fMRI datasets in Appendix Figure 1. We only plot the pairwise correlation values found to be significant for each individual dataset (one-sample t-test, FDR corrected for multiple comparisons across ROI pairs at alpha level 0.05 [2]). For both datasets, at the group and participant level we find that the amount of functional connectivity between language ROIs varies but the vast majority of ROI pairs have significant correlations. As observed in Figure 3.2 from the simulations, both high functional connectivity and high encoding model performance can be caused by multiple settings of the underlying shared information between the brain sources, stimulus, and stimulus representations. Since the overwhelming majority of ROI pairs have significant correlations and the individual ROI have significant encoding model performance, it is still not possible to disentangle the different cases even when combining these two metrics.

Source generalization. We further investigate the source generalization to understand if the differences in the amount of shared information between language ROIs are due to process-

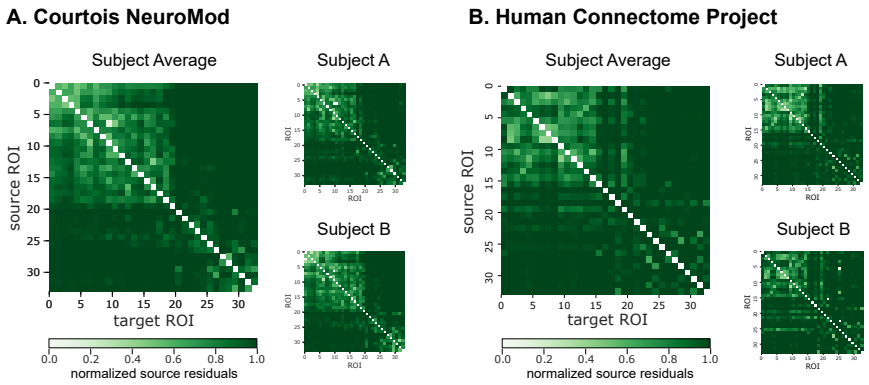

Figure 3.5: Source Residuals. ROI pairs with high norm. source residuals (dark green) are processing unique information related to the stimulus representations. These ROI pairs with high norm. source residuals are consistent at the group and individual level in both datasets.

ing shared information related to the stimulus representations. We present the pairwise source generalization for the 33 language ROIs in both fMRI datasets in Figure 3.4. The source generalizations are normalized by the ISC, as an estimate of the "noise ceiling" (see Appendix for more details and suggestions for other types of normalization when investigating different scientific questions). In both datasets, at the group and participant level we find that there are differences in pairwise language ROI source generalizations. However, it is unclear if these differences are due to true differences among language ROIs in the amount of shared information related to the stimulus or to a limitation of using an ELMo embedding as our stimulus representation. Source generalization alone cannot distinguish if the relationship between two ROI is case B or C from Figure 3.1 with respect to the stimulus.

Source residuals. Next we investigate the second proposed metric, source residuals to understand the amount of unique stimulus-related information processed by language ROIs. We present the pairwise source residuals for the 33 language ROIs in both fMRI datasets in Figure 3.5. The source residuals are normalized by the ISC, an estimate of the "noise ceiling". In both datasets, we find differences in the source residuals at the group and participant level. The high source residuals reveal that the majority of regions process some unique information about the stimulus that cannot be fully accounted for by any of the other considered regions. However, this does not mean that the regions do not also process some shared information about the stimulus, which would have been removed during the residual computation (i.e. there will be high source residuals in both cases A and C in Fig. 3.1.

Source generalization and source residuals. We reexamine the case study presented in Section 1. As we are interested in answering whether language ROIs are indeed processing the same information about the stimulus or whether we cannot differentiate among the ROI due to methodological limitations, we focus on six bilateral language ROIs that have been previously shown to be difficult to disentangle [5, 50]. We present an example using the proposed framework to infer each of the three relationship cases (e.g. A, B, or C) within the six ROIs in Figure 3.6. These relationships are consistent across both datasets. The relationships between ROI pairs can

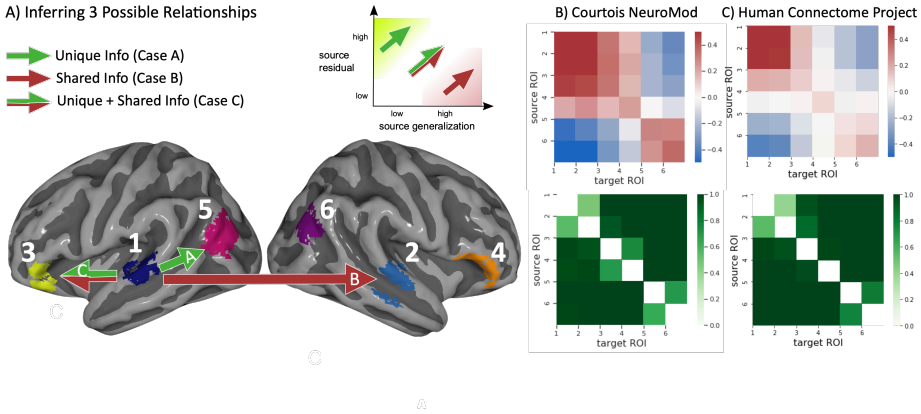

Figure 3.6: Source Generalization and Source Residuals. We use the proposed framework to infer an example of each of the three relationships between two brain sources, stimuli and stimuli representations.

be asymmetric (ie. ROI A could better generalize to ROI B than ROI B to ROI A), therefore we depict the relationships as a directed edges from the source ROI to the target ROI. The inferred relationship, case A, between ROI 1 and ROI 5 suggests that it is possible to differentiate the information processed between two ROI with significant encoding model performance. The inferred relationship, case B, between ROI 1 and ROI 2 support previous findings that bilateral language ROIs process shared information with respect to the stimulus. The inferred relationship, case C, between ROI 1 and ROI 3 shows that two ROI can process both shared and unique information. In this case the stimulus representation can limit our understanding of the amount of shared versus unique information processed between the two ROI.

3.6 Discussion

In this work, we conceptually breakdown the possible underlying relationships between two brain sources and the presented stimuli where both sources are predicted by the same feature space and present two metrics to distinguish them. We next present simulations that show limitations of commonly used methods for disambiguating these relationships. Finally, we showcase the use of these metrics in two fMRI datasets with naturalistic stimuli showing when we can disambiguate and when we are limited by the stimulus representation. Our results generalize across these two datasets which capture different populations and are acquired by different labs in different countries with very different experimental setups and scanning parameters.

In contrast to the encoding model performance, our proposed metrics disentangle the three cases outlined in Fig.3.1 A-C. Note that in case C both source residuals and source generalization are high which indicates that the two brain sources process shared and unique information about the stimulus. We expect that a fully descriptive stimulus representation will capture this unique information that is shared between the stimulus and each individual brain source, so if both metrics are high we conclude that the stimulus representation used in the encoding model is not informative enough to disentangle the information processed in the two brain sources. This allows us

to infer that we in fact need better stimulus representations rather than that the two brain sources are processing identical information about the stimulus. This interpretation suggests that Region 1 (middle superior temporal gyrus) in Fig. 3.6 and Region 3 (left inferior frontal gyrus) may be more easily distinguishable in the future using an encoding model if we have new stimulus representations that capture unique information that either region is processing about the stimuli. Our framework can be used as a test for future representations–if these future representations lead to a high encoding model performance and the source residuals between Region 1 and Region 3 continue to be high in the investigated stimulus set, but the source generalization using the new feature space decreases, then the new stimulus representation better captures some of the unique information processed by at least one of the regions.

One limitation of our proposed source residual metric is that the residuals contain information about the unique information in both regions that were used to compute the residuals. Isolating the residual information in an individual region may further improve our ability to disambiguate the information processed by different regions. We hope that our approach can serve as basis for such future work. Overall, our proposed framework is a promising new tool for computational neuroscientists who are interested in mapping information processing in the brain.

Chapter 4

Augmenting encoding models with nonlinearity

4.1 Introduction

So far, we have focused our analysis on the most frequently used computational tool for brain mapping - the linear encoding model. Most studies have relied on linear encoding models because they are seen as being more interpretable and data-efficient compared to their nonlinear counterparts. However, as discussed in Section 2.1, there is a growing interest in aligning brain activity with complex representations derived from deep learning models that seem to capture something generic about the stimulus. Because we do not know exactly what kinds of stimulus-related information is captured by these network-derived representations, it is difficult for any encoding approach we can come up with, whether linear or nonlinear, to be accompanied by the promise of interpretability that we enjoyed when dealing with simpler, hand-crafted stimulus representations. In addition, some of the practical difficulties previously associated with training a nonlinear encoding model have now become surmountable with access to larger-scale datasets and more powerful computational resources.

Nonlinear encoding models can allow us to align stimulus-related information that may be nonlinearly encoded within the NLP and fMRI representations. If at all present, we would not be able to access and align such information by restricting ourselves to linear encoding models. Nonlinear models also lend themselves well to situations where the same pockets of information may be relevant to predicting a large group of voxels. In this chapter, we look at how model complexity can affect encoding performance by comparing the widely used linear encoding model with a multi-layer nonlinear encoding model.

4.2 Related Work

Prior work has revealed a great deal of overlap between what state-of-the-art NLP models and brain activity recordings capture about the same language input. Most of this work has relied on studying linear mappings from the NLP representations to brain activity [61], [37], [54], [60].

More recently, we have started to see more work showing the potential benefits nonlinearity can bring to brain mapping efforts. Recent work has shown that we can model nonlinear interactions between fMRI responses from different brain regions to achieve better performance than contemporary linear approaches [1]. Researchers in neuroscience have started to delve deeper into the debate we address here by providing a thorough examination of for and against why future research efforts should focus more on exploiting linear encoding models rather than exploring nonlinear ones [36].

4.3 Methods

We align fMRI recordings of participants reading naturalistic text word-by-word with intermediate representations from BERT [11], a state-of-the-art NLP model, generated by passing in the same text.

fMRI data We use fMRI recordings of 8 healthy participants reading chapter 9 of Harry Potter and the Sorcerer's Stone [52], provided by Wehbe et. al [66]. For more details, refer to *fMRI data: Harry Potter and the Sorcerer's Stone* in Section 1.1.

BERT data Let n denote the total number of words presented and t denote the total number of TR intervals that span the presentation of this text. We generate the corresponding BERT representation x_i for word w_i, where $i \in [0, n)$, by passing the 10 most recent words $(w_i, w_{i-1}, \ldots, w_{i-9})$ to BERT and collecting the layer 8 representations from the network. Each x_i is then reduced from a 768-dimensional vector to 10 dimensions via Principal Components Analysis (PCA) [18] that has been fit on the training data. This specific configuration of context length and layer depth was chosen as it has previously been shown to align well with the same brain data [60]. At each time point, we compute a single representation z_j, where $j \in [0, t)$, by averaging the x_i's corresponding to the words that appear in that interval.

Encoding models We estimate 3 distinct encoding models that take the preprocessed BERT representations as input and predict the fMRI voxel activities associated with reading the same set of words:

- *Linear-Analytical*: a linear model with weights estimated using a closed-form solution
- *Linear-GD*: a linear model trained using gradient descent
- *MLP-GD*: a 1-hidden layer multi-layer perceptron (MLP) trained using gradient descent

As mentioned earlier, previous studies have primarily relied on using the *Linear-Analytical* approach to align NLP representations with brain activity. We provide details on model training under this approach in section 1.1. Figure 4.1 depicts the two types of architectures we use in this study. Adding a shared hidden layer between the inputs and outputs allows us to benefit from two key features that separate the multi-layer architecture from the single-layer linear architecture: (i) the introduction of nonlinearity in the form of a ReLU activation function at this intermediate layer and (ii) the ability to predict all voxel activities jointly using a shared hidden layer.

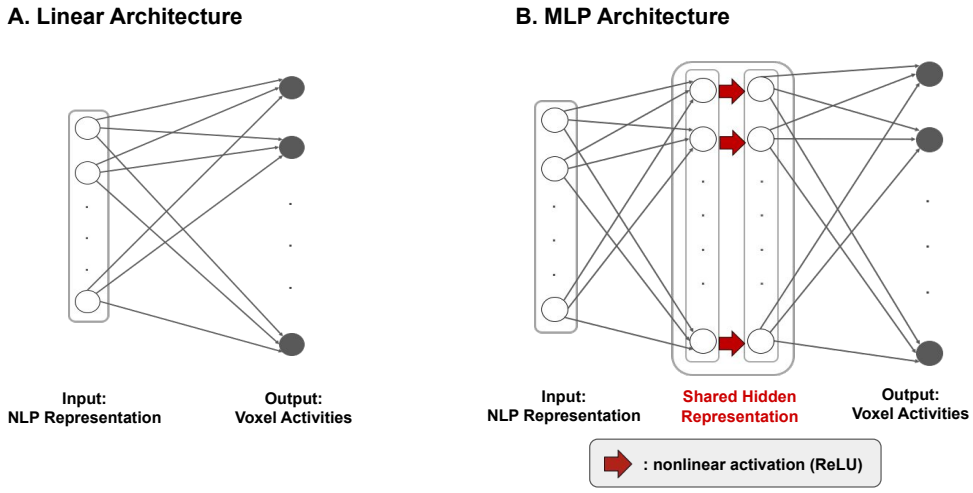

Figure 4.1: A side-by-side comparison of the single-layer linear architecture used in *Linear-Analytical* and *Linear-GD* with the multi-layer nonlinear architecture used in *MLP-GD*.

However, the introduction of nonlinearity in our encoding model comes at a price as we can no longer use a closed-form solution to estimate its weights. We instead use gradient descent, an iterative optimization algorithm, when modeling fMRI-BERT alignment as a nonlinear function. As mentioned earlier, we call this approach *MLP-GD*. Note that this added difference in how the weights of *Linear-Analytical* and *MLP-GD* are estimated can make it difficult to distinguish performance differences related to the training method from performance differences that may be attributed to nonlinearity and a shared hidden layer. We therefore include *Linear-GD* as an intermediate point of comparison between the two approaches to make this distinction recognizable in the experiments that follow.

Evaluation We evaluate encoding performance by computing the mean Pearson correlations between the models' predictions and the true voxel activities in held-out data for regions of interest (ROI) known to be consistently activated during language processing [15]. We also compare these correlations against a noise ceiling estimated using a method put forth by [54]. We compute pairwise correlations between participants' fMRI recordings and use them to obtain a conservative estimate of the ceiling value at each voxel. This allows us to estimate an upper bound on the amount of brain signal that the best encoding model can hope to explain. We perform paired sample t-tests, where each pair corresponds to an individual participant, to test for significant differences between our models' performances and to compare each of them to the computed noise ceiling. We report one-sided p-values on each ROI after using false discovery rate to control for multiple comparisons [2] (significance level 0.05). We also use pycortex [21] to visualize voxel correlations on a 3D brain surface.

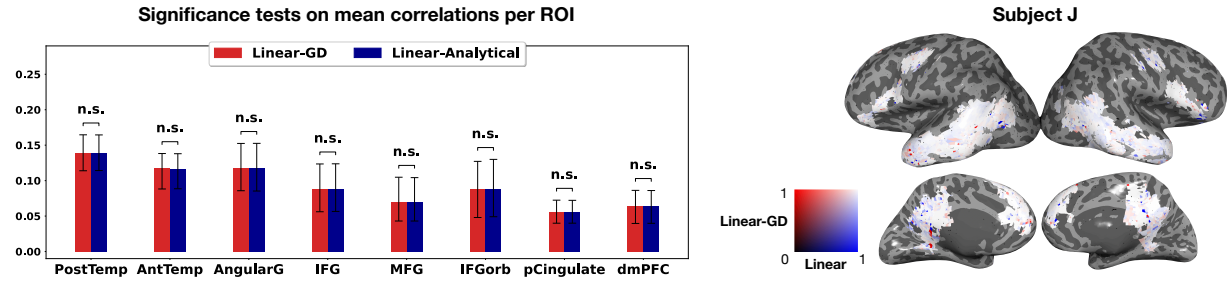

Figure 4.2: We compare *Linear-Analytical* with *Linear-GD*. The plot on the left shows mean correlations recorded across 8 participants at the ROI-level. *n.s.* indicates that the difference between the two models' encoding performances was not found to be significant in that ROI. On the right, we present a cortical visualization of this comparison on an individual subject (subject J). Here, white indicates that both models predict the specified voxels well.

4.4 Results

***Linear-Analytical* and *Linear-GD* have similar performance** We first investigate whether the shift in training method from a closed-form solution to gradient descent leads to a difference in performance by comparing *Linear-Analytical* with *Linear-GD*. We do not observe a significant difference on the whole ROI level here ($p > 0.5$ for all ROI). Figure 4.2 shows the per-ROI mean correlations across subjects from both models and a voxel-level visualization of the comparison between the two normalized correlations for an individual participant. We find that it is possible to replace the analytical approach with gradient descent in this setting without sacrificing encoding performance. All cortical visualizations we present here are on an individual subject J. We show the same comparison for the remaining 7 participants in section A.7 of the appendix.

Room for improvement in *Linear-GD* We further investigate how the linear encoding model's performance compares with our noise ceiling estimates. In Figure 4.3, we again present our results from significance testing on whole ROI and visualize normalized correlations at the voxel-level for the same participant as before. We observe that *Linear-GD* performs on par with the noise ceiling in 5 ROI ($p > 0.2$). Differences in the remaining 3 ROI - the dorsomedial prefrontal cortex (dmPFC, $p = 0.009$), inferior frontal gyrus pars orbitalis (IFGorb, $p = 0.05$) and posterior cingulate (pCingulate, $p = 0.06$) - approach significance levels. The corresponding cortical visualizations of differences between *Linear-GD* and the noise ceiling seem to further confirm that there is room for improvement in encoding performance at these ROI.

***MLP-GD* shows promising results in some of these improvement areas** Next, we compare the encoding performance of *Linear-GD* and *MLP-GD*. In Figure 4.4, we present the per-ROI mean correlations across subjects from both models along with our qualitative results from this comparison. Differences between *Linear-GD* and *MLP-GD* evaluated based on the whole ROI are not found to be significant ($p > 0.05$). However, our voxel-level visualization of this comparison shows that for a majority subsample of the participants, *MLP-GD* seems to be outperforming

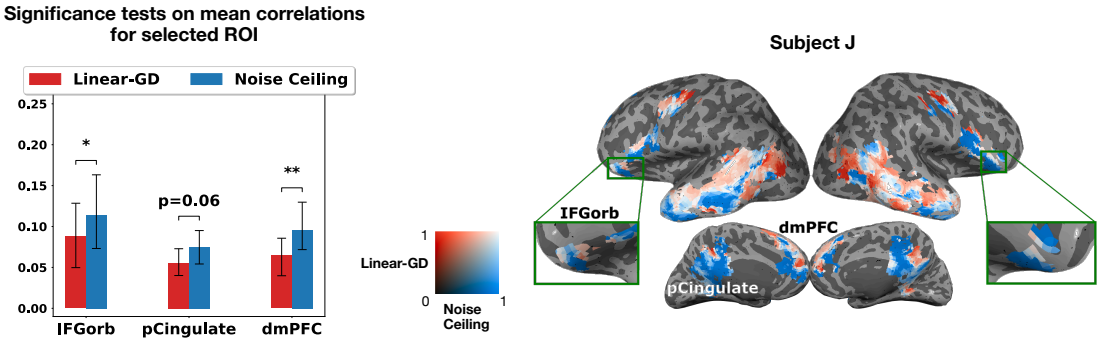

Figure 4.3: We compare the encoding performance of *Linear-GD* with the noise ceiling. The plot on the left shows mean correlations recorded across 8 participants for selected ROI where we observe the most difference. * ($p \leq 0.05$) and ** ($p \leq 0.01$) indicate where our paired t-tests show a significant difference between the two quantities and *n.s.* indicates where a significant difference was not found. On the right, we present a cortical visualization of this comparison on the same subject as before (subject J). Here, blue indicates voxels where *Linear-GD* is unable to match the noise ceiling and white indicates where the two are relatively similar. We expand on the significance of red voxels in Figure 4.5.

Linear-GD in specific clusters of voxels in the dmPFC, IFGorb and pCingulate. However, we also observe that our version of *MLP-GD* is sub-optimal given that its performance was surpassed by the less expressive *Linear-GD*. Although the same training method is used for both models, optimizing a nonlinear model can pose difficulties as it requires solving a non-convex problem and we have no reference point in hand (like a closed-form solution) to identify whether the setting of weights we settle at is still far from the true optimum.

Room for improvement in noise ceiling Figure 4.5 helps bring attention to another qualitative observation that suggests that the noise ceiling estimates we compute for our data are sub-optimal. We find that the performance of *Linear-GD* is sometimes able to exceed the estimated noise ceiling, particularly for clusters of voxels within the posterior temporal lobe (PostTemp), angular gyrus (AngularG) and middle frontal gyrus (MFG).

4.5 Discussion

This study aimed to investigate how varying the complexity of an encoding model can affect its ability to predict fMRI data of participants reading naturalistic text. We compared different approaches to encoding within this setting based on linear and nonlinear, multi-layer network architectures. We began by verifying that using gradient descent over a closed-form solution does not significantly alter the resulting encoding performance. We then used noise ceiling estimates at the voxel-level, computed based on pairwise correlations between participants' fMRI recordings, to show that there is room for improvement in the predictions we obtain from a linear encoding model, particularly in 3 ROI - dmPFC, IFGorb and pCingulate. Interestingly, our qualitative analysis revealed clusters within these ROI that were consistently better predicted by a

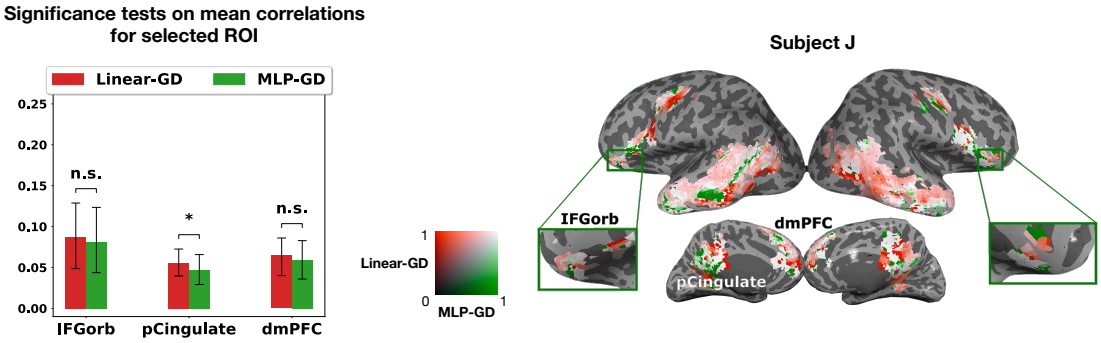

Figure 4.4: We compare the encoding performance of *Linear-GD* with that of MLP-GD. The plot on the left shows mean correlations recorded across 8 participants for selected ROI where we observe the most difference. * ($p \leq 0.05$) indicates where our paired t-tests show a significant difference between the two quantities and *n.s.* indicates where a significant difference was not found. On the right, we present a cortical visualization of this comparison on the same subject as before (subject J). Here, green indicates voxels where *Linear-GD* is unable to match *MLP-GD*, red indicates where *Linear-GD* outperforms *MLP-GD* and white indicates where the two are relatively similar.

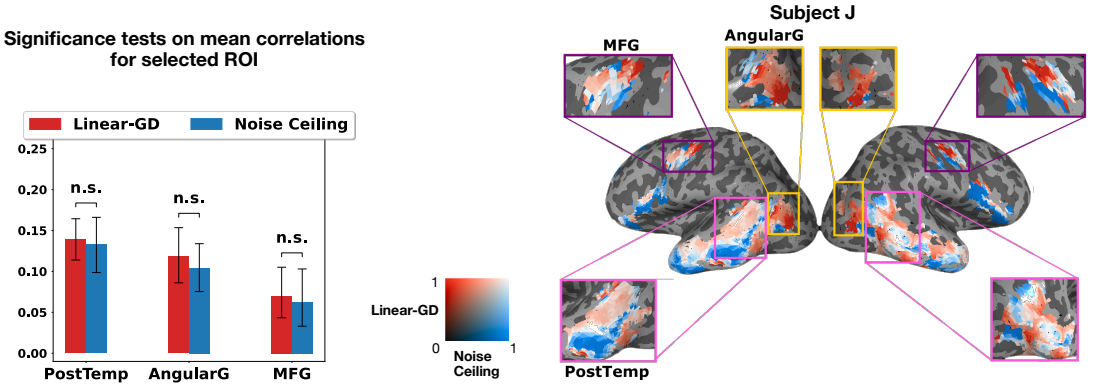

Figure 4.5: Similar to Figure 4.3, we again compare the encoding performance of *Linear-GD* with the noise ceiling but emphasize a different phenomenon here. The plot on the left shows mean correlations recorded across 8 participants for selected ROI where we find that *Linear-GD* seems to be performing better the upper bound established by our ceiling estimates. Although a significant difference was not found in these ROI, we find clusters of voxels (colored red) within them where our noise ceiling is suboptimal.

nonlinear model that we trained on the same data. One interpretation of this finding is that these clusters may process different information from the rest of the region – information that only a nonlinear alignment can reveal – but further investigation is necessary.

Future work Although we observe clusters in specific brain regions where the linear encoding model seems to surpass the noise ceiling, this work shows that there might be gaps in the linear architecture that we can fill with better ceiling estimates. A better noise ceiling may provide stronger evidence for our observations and highlight other areas where the encoding model can be improved upon as a guide to future research. We also found clusters where our nonlinear encoding model was unable to match its linear counterpart's performance, despite its greater expressive power. It is possible that the limited availability of brain recordings made it difficult to effectively optimize our nonlinear encoding model. Performing experiments with larger datasets, more nonlinear model architectures and more hyperparameters during training is therefore another avenue for further exploration. This can help bring us closer to the globally optimal setting of model weights so we may fully harness the superior expressive power that a larger model can offer.

Chapter 5

Conclusion

Encoding models have allowed us to study language processing in the human brain by leveraging specific stimulus-related information. In this thesis, we revisit the encoding model framework to see how it can better support complex representations of natural stimuli. We start by showing that the existing framework does not always let us disentangle where specific stimulus-related information is processed in the brain, especially when working with naturalistic stimuli like movies. As an initial step towards overcoming this limitation, we propose a promising new encoding model framework that can allow us to recognize groups of brain regions where similar stimulus-related information is processed. Finally, we show a proof of concept that derestricting the encoding model by allowing it to learn nonlinear brain mappings might let us align more stimulus-related information with brain measurements than a linear model. Overall, this work puts forth preliminary evidence for two promising ways in which the current encoding model framework can be modified so we may better adapt to increasingly complex representations of naturalistic stimuli.

Bibliography

[1] Stefano Anzellotti, Evelina Fedorenko, Alfonso Caramazza, and Rebecca Saxe. Measuring and modeling transformations of information between brain regions with fmri. *bioRxiv*, 10/2016 2016. 4.2

[2] 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. 2.3, 3.5, 4.3

[3] 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. 3.5

[4] 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. 3.2

[5] Charlotte Caucheteux, Alexandre Gramfort, and Jean-Remi King. Decomposing lexical and compositional syntax and semantics with deep language models. *arXiv preprint arXiv:2103.01620*, 2021. 3.1, 3.2, 3.5, 3.5

[6] Charlotte Caucheteux, Alexandre Gramfort, and Jean-Remi King. Gpt-2's activations predict the degree of semantic comprehension in the human brain. *bioRxiv*, 2021. 1

[7] 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. 1

[8] 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`. 3.2

[9] Wendy A de Heer, Alexander G Huth, Thomas L Griffiths, Jack L Gallant, and Frédéric E Theunissen. The hierarchical cortical organization of human speech processing. *Journal of Neuroscience*, 37(27):6539–6557, 2017. 3.2

[10] 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. 3.1, 3.2

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. 4.3

[12] 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. 3.5

[13] Oscar Esteban, Ross Blair, Christopher J. Markiewicz, Shoshana L. Berleant, Craig Moodie, Feilong Ma, Ayse Ilkay Isik, Asier Erramuzpe, Mathias Kent, James D. and-Goncalves, Elizabeth DuPre, Kevin R. Sitek, Daniel E. P. Gomez, Daniel J. Lurie, Zhifang Ye, Russell A. Poldrack, and Krzysztof J. Gorgolewski. fmriprep. *Software*, 2018. doi: 10.5281/zenodo.852659. A.5.2

[14] 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*, 2018. doi: 10.1038/s41592-018-0235-4. A.5.2

[15] 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/. 3.1, 3.5, 4.3

[16] 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. 3.5

[17] David Freedman, Robert Pisani, and Roger Purves. Statistics (international student edition). *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*, 2007. 2.3

[18] Karl Pearson F.R.S. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. doi: 10.1080/14786440109462720. 4.3

[19] Alona Fyshe. Studying language in context using the temporal generalization method. *Philosophical Transactions of the Royal Society B*, 375(1791):20180531, 2020. 3.2

[20] 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. 3.2

[21] 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. (document), 2.3, 4.3, 6

[22] 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. 3.5

[23] 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. 2.2, 3.5

[24] 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. 1.1

[25] 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. 1

[26] 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. 3.5

[27] 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/. 1.1

[28] 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. 1

[29] 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/. 3.2

[30] Stefan Haufe, Frank Meinecke, Kai Görgen, Sven Dähne, John-Dylan Haynes, Benjamin Blankertz, and Felix Bießmann. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage*, 87:96–110, 2014. 3.2

[31] 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. 3.2

[32] 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. A.3

[33] 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`. 1, 3.2

[34] 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. 1, 1.1, 3.1, 3.2, 3.5

[35] Alexander G. Huth, Wendy A. De Heer, Thomas L. Griffiths, Frédéric E. Theunissen, and Jack L. Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, 4 2016. ISSN 14764687. doi: 10.1038/nature17637. A.5.3

[36] Anna Ivanova, Martin Schrimpf, Leyla Isik, Stefano Anzellotti, Noga Zaslavsky, and Evelina Fedorenko. Is it that simple? the use of linear models in cognitive neuroscience. 2020. 4.2

[37] Shailee Jain and Alexander Huth. Incorporating context into language encoding models for fmri. In *Advances in neural information processing systems*, pages 6628–6637, 2018. 1, 4.2

[38] 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. A.5.3

[39] Kendrick N Kay, Thomas Naselaris, Ryan J Prenger, and Jack L Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352, 2008. 1

[40] 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. 3.2

[41] Mark D Lescroart and Jack L Gallant. Human scene-selective areas represent 3d configurations of surfaces. *Neuron*, 101(1):178–192, 2019. 3.2, A.3

[42] Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew E Peters, and Noah A Smith. Linguistic knowledge and transferability of contextual representations. *arXiv preprint arXiv:1903.08855*, 2019. 1

[43] 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. 1

[44] Thomas Naselaris, Kendrick N. Kay, Shinji Nishimoto, and Jack L. Gallant. Encoding and decoding in fMRI, 5 2011. ISSN 10538119. 3.2

[45] Samuel A Nastase, Ariel Goldstein, and Uri Hasson. Keep it real: rethinking the primacy of experimental control in cognitive neuroscience. *NeuroImage*, 222:117254, 2020. 1

[46] 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. 1, 1.1

[47] 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*, 21(19):1641–1646, 10 2011. ISSN 09609822. doi: 10.1016/j.cub.2011.08.031. URL /pmc/articles/PMC3326357/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3326357/. A.5.3

[48] Anwar O Nunez-Elizalde, Alexander G Huth, and Jack L Gallant. Voxelwise encoding models with non-spherical multivariate normal priors. *Neuroimage*, 197:482–492, 2019. 3.2

[49] 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. 2.1, 3.1, 3.5

[50] Aniketh Janardhan Reddy and Leila Wehbe. Syntactic representations in the human brain: beyond effort-based metrics. *bioRxiv*, 2020. 3.1, 3.2, 3.5, 3.5

[51] 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. 3.5

[52] J. K Rowling. *Harry Potter and the sorcerer's stone*. A.A. Levine Books, New York, 1st american ed edition, 1998. ISBN 0590353403. 1.1, 4.3

[53] 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/. 1.1

[54] Martin Schrimpf, Idan A Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy G Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. Artificial neural networks accurately predict language processing in the brain. *BioRxiv*, 2020. 1, 4.2, 4.3

[55] 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/. 3.5

[56] Erez Simony, Christopher J. Honey, Janice Chen, Olga Lositsky, Yaara Yeshurun, Ami Wiesel, and Uri Hasson. Dynamic reconfiguration of the default mode network during narrative comprehension. *Nature Communications*, 7(1):1–13, 7 2016. ISSN 20411723. doi: 10.1038/ncomms12141. URL https://www.nature.com/articles/ncomms12141. 3.2

[57] Saurabh Sonkusare, Michael Breakspear, and Christine Guo. Naturalistic stimuli in neuroscience: critically acclaimed. *Trends in cognitive sciences*, 23(8):699–714, 2019. 1

[58] 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. 1.1

[59] Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*, 2019. 1

[60] 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. 1, 1.1, 2.2, 3.5, 3.5, 4.2, 4.3

[61] Mariya Toneva, Tom M Mitchell, and Leila Wehbe. Combining computational controls with natural text reveals new aspects of meaning composition. *bioRxiv*, 2020. 2.2, 3.2, 3.5, 3.5, 4.2, A.3

[62] 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, 2010. 3.2

[63] 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. 1.1

[64] Aria Y Wang, Leila Wehbe, and Michael Tarr. Neural taskonomy: Inferring the similarity of task-derived representations from brain activity. 2019. 1

[65] 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 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0112575. URL https://dx.plos.org/10.1371/journal.pone.0112575. 1, 1.1, 3.1, A.5.3

[66] 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, nov 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0112575. URL http://dx.plos.org/10.1371/journal.pone.0112575. 1.1, 4.3

[67] 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, 2014. 1, 1.1

[68] Leila Wehbe, Aaditya Ramdas, Rebecca C Steorts, Cosma Rohilla Shalizi, et al. Regularized brain reading with shrinkage and smoothing. *Annals of Applied Statistics*, 9(4): 1997–2022, 2015. 3.2

[69] Leila Wehbe, Idan A 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. *bioRxiv*, 2020. A.3

[70] 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. 3.2

[71] Michael C-K Wu, Stephen V David, and Jack L Gallant. Complete functional characterization of sensory neurons by system identification. *Annu. Rev. Neurosci.*, 29:477–505, 2006. 3.2

[72] 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. 1

Appendix

A.1 Functional Connectivity Results

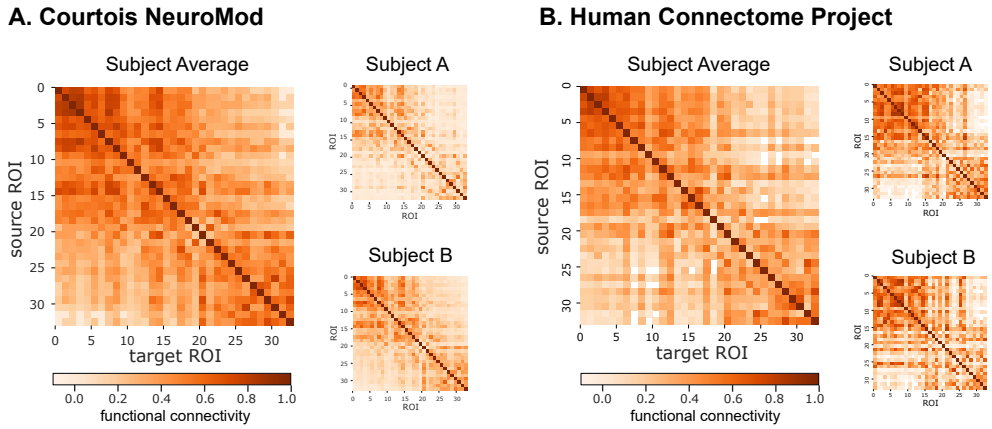

Figure 1: Functional Connectivity. Significant pairwise correlations of the 33 language ROIs (corrected at level 0.05). The overwhelming majority of ROI pairs have significant correlations.

A.2 Relationship of Special Cases A-C in Figure 1 to Most General Case

In Figure 2, we present a Venn diagram that captures all possible relationships among two brain measurements, the presented stimulus, and the stimulus representation. All possible cases can be obtained by varying the amount of information that makes up each of the regions annotated with a red number, as well as the analogous regions in the blue brain source 1. In the main paper, we focus on three specific cases, presented in Figure 1 and replicated at the bottom of Figure 2 for ease of visualization. These cases were selected because they are all possible candidates for the underlying relationships that lead to the real-world example of the inference problem described in Section 1, because an encoding model would perform equally well at predicting both sources in each of these 3 cases.

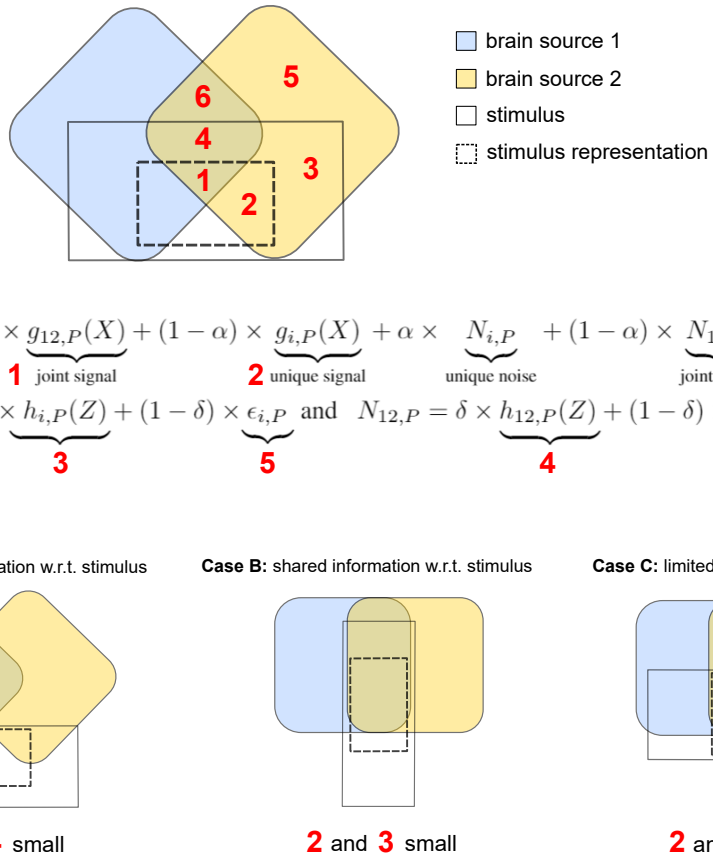

Figure 2: (Top) Most general Venn diagram that captures all possible underlying relationships between two brain measurements, the presented stimulus, and the stimulus representation. (Middle) Annotated data generation model in Eq. 7. (Bottom) Special cases considered in the main paper, that we argue cannot be disambiguated solely through encoding model performance.

A.3 Metric Normalizations

The main metrics of interest defined in Section 3 are encoding model performance, source generalization, and source residuals. In the simple setting where all annotated regions in Figure 2 are independent of each other, the encoding model performance is proportional to annotated regions **1 + 2**, source generalization to **1**, and source residuals to **2 + 3** (and to the analogous **2 + 3** regions on the blue brain source side). For some scientific questions, it may be more informative to normalize these metrics in different ways. For example, one may normalize the source generalization by the encoding model performance to compute the proportion of $\frac{1}{1+2}$ (i.e. the proportion of information shared between a brain source and the stimulus representation that is also shared by a second brain source). This metric is identical to the one proposed by Toneva et al. [61]. Another type of normalization that we find informative in the current work is the intersubject correlation (ISC), which is proportional to **1 + 2 + 3 + 4** (i.e. the information shared between a brain source and the stimulus). This metric can be thought of as an estimate of the maximum possible performance (i.e. the noise ceiling). A similar metric was used as an estimate of the

noise ceiling by Wehbe et al. [69], though the authors did not make the connection to ISC explicitly. Note that the ISC across a dataset of more than two subjects is computed as the average of the pairwise ISC (i.e. the ISC for 1 of 6 subjects is the average across the ISC computed between that subject and the remaining 5 subjects). Following previous work [32, 41], we normalize all of our metrics by the square-root of the noise-ceiling, yielding normalized correlation values.

We hope that the conceptual breakdown of the different possible relations that we present in Figure 1 and Figure 2 will help other researchers choose the most relevant normalization for their questions of interest.

A.4 Additional Simulation Results

We present how encoding model performance, functional connectivity, source generalization and source residuals vary as we allow $\alpha, \delta \in [0, 1]$ to vary with respect to each other in Eq. 3.6. Recall that α lets us control how much of the information in both brain sources that is related to the stimulus representation is shared between the two brain sources. Since Case A is characterized by unique information that is related to the stimulus representation and Cases B & C are characterized by shared information related to the stimulus representation, increasing α from 0.0 to 1.0 means the simulated data departs from Case A and more closely resembles Cases B & C. Also recall that given a high value of α, varying δ lets us control the amount of information captured by both brain sources that is related to the stimulus but absent in the stimulus representation. So as δ is increased from 0.0 to 1.0 at high α, the simulated data more closely resembles Case C than Case B.

Fig. 3 shows that source generalization is the only metric of the four we consider that can be used to separate Case A from Cases B & C - i.e., distinguish between brain source data simulated using low and high values of α respectively at any fixed δ choice. This aligns with our observations from Fig. 3.2A, where we concluded that looking at source generalization can help with disentangling Case A from Cases B & C in our simulations. If we identify that we are not in Case A, we may further narrow down which case we are in using any metric that lets us distinguish between brain source data simulated at different values of δ when α is high. Fig. 3 shows that at high α, source residual increases with δ, and therefore, can be useful when separating Case B from Case C. This also aligns with our observations from Fig. 3.2B.

A.4.1 Varying signal-to-noise ratio in simulated brain source data

To test the boundaries of what each metric can tell us about the information captured by a pair of brain sources, we further extend Eq. 3.6 to allow for variation in the amount of stimulus representation related information that each simulated brain source captures as follows:

$$Y_{i,P} = \beta_i \underbrace{[\alpha \times g_{12,P}(X) + (1-\alpha) \times g_{i,P}(X)]}_{\text{signal}} + (1-\beta_i) \underbrace{[\alpha \times N_{i,P} + (1-\alpha) \times N_{12,P}]}_{\text{noise}} \quad (1)$$

where $N_{i,P} = \delta \times h_{i,P}(Z) + (1-\delta) \times \epsilon_{i,P}$ and $N_{12,P} = \delta \times h_{12,P}(Z) + (1-\delta) \times \epsilon_{12,P}$.

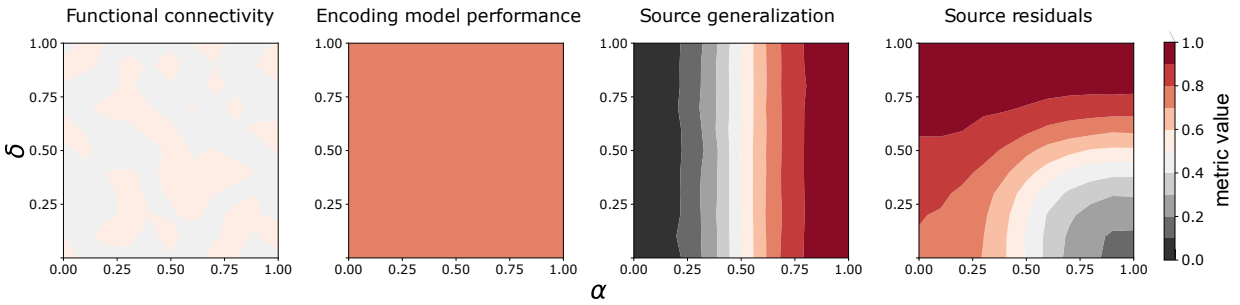

Figure 3: Plotting how each metric varies under simulations performed at different settings of α, δ.

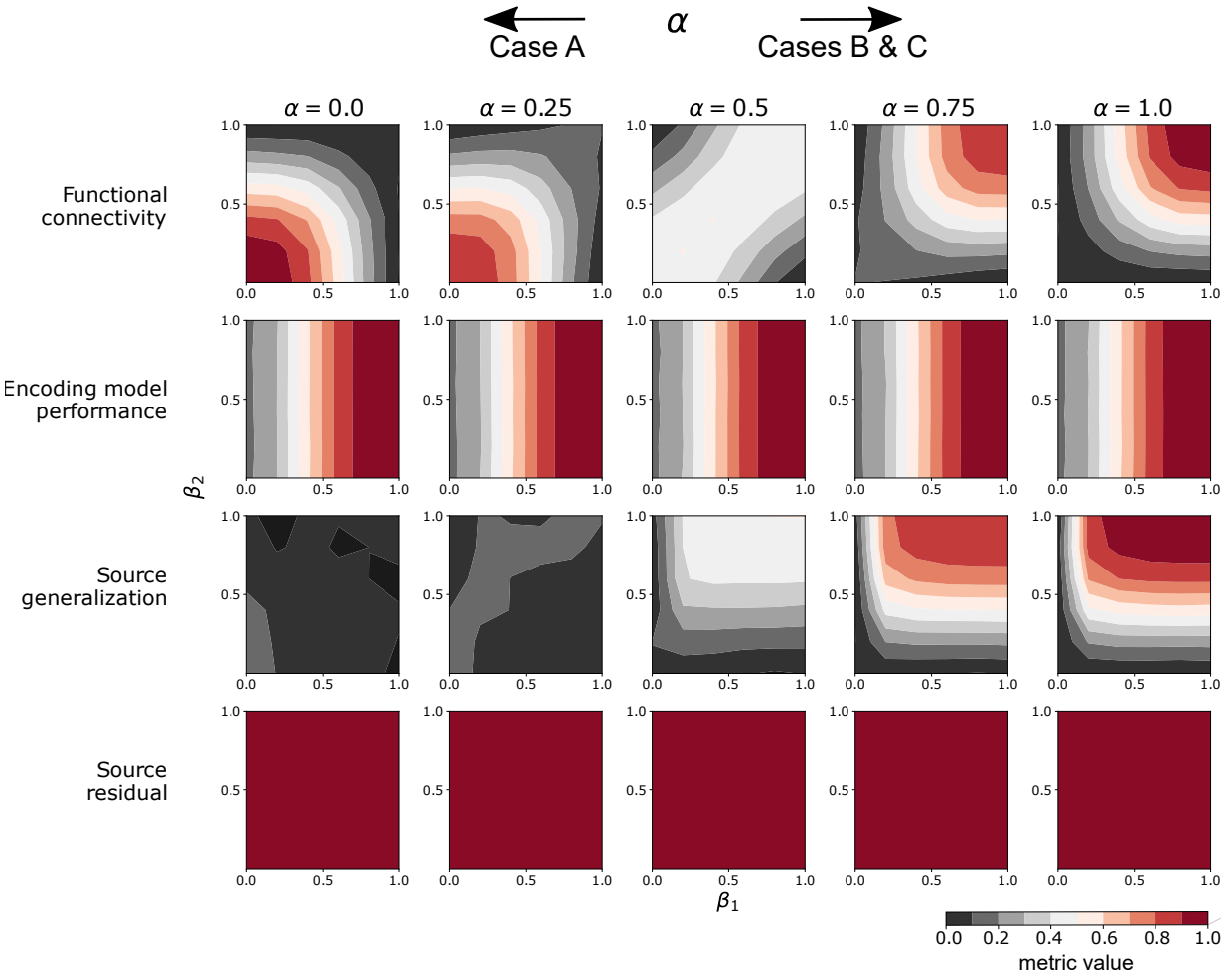

Figure 4: Extending on Fig. 3.2A, this figure shows how each metric varies under simulations performed at different signal-to-noise ratios as we vary α, β_1, β_2 when $\delta = 1.0$ is fixed.

Here, we introduce an additional type of parameter $\beta_i \in [0, 1]$, that is used to adjust the amount of stimulus representation related information that is captured by the ith brain source. In this context, the stimulus representation related information can be viewed as the signal that is retrievable by an encoding model. The remaining information, whether related to the stimulus or not, can be viewed as the noise that an encoding model that only has access to the stimulus representation cannot explain.

First, we focus on the separation between Case A and Cases B & C by varying α. In Fig. 4, we show how each metric varies as we control the signal-to-noise ratio in both simulated brain sources by varying β_1 and β_2 for each setting of α. We fix $\delta = 1.0$ here, but similar trends can be observed for other choices of fixed $\delta \in [0, 1]$ as well. We observe that encoding model performance and source residuals cannot allow us to distinguish between different α values. Note also that from looking at functional connectivity, one cannot separate when both brain sources mostly capture shared noise (low β_i's, low α) from when they mostly capture shared signal (high β_i's, high α). We observe that given sufficient signal in both brain sources ($\beta_1, \beta_2 \gg 0$), source generalization increases as α increases. However, under conditions of little to no signal, source generalization cannot be used to distinguish between different α values. These results suggest that source generalization is a useful metric to separate Case A (low α) from Cases B & C (high α) only when the encoding models that output the predictions used to compute this metric are able to perform relatively well on the brain sources they are trained on.

Next, we consider the separation between Case B and Case C by maintaining a high α and varying δ. In Fig. 5, we show how each metric varies as we control the signal-to-noise ratio in each brain source by varying β_1 and β_2 for each setting of δ. We fix $\alpha = 1.0$ here as we've seen that a low value for α would generate brain data that more closely resembles Case A. We find that encoding model performance, functional connectivity and source generalization are not useful to distinguish between different values of δ in our simulations. We observe that given sufficient noise in brain source 1 ($\beta_1 \ll 1$), the source residual increases as δ increases. However, when most or all of the information captured by brain source 1 is related to the stimulus representation, the source residual for this brain source cannot be used to distinguish between different δ values. These results suggest that the source residual can help us separate Case B (low δ) from Case C (high δ) only when there is some information captured by the source that cannot be explained by an encoding model.

A.5 Data Preprocessing

A.5.1 HCP

Our analyses are performed with the 3105 TRs (51 minutes and 45 seconds) suggested for analysis in the HCP documentation. These exclude rest periods and the first 6 TRs of each movie clip within a movie run. Individual-level results are presented for the six participants with the highest encoding model performance averaged over the 55 Shen atlas language ROIs.

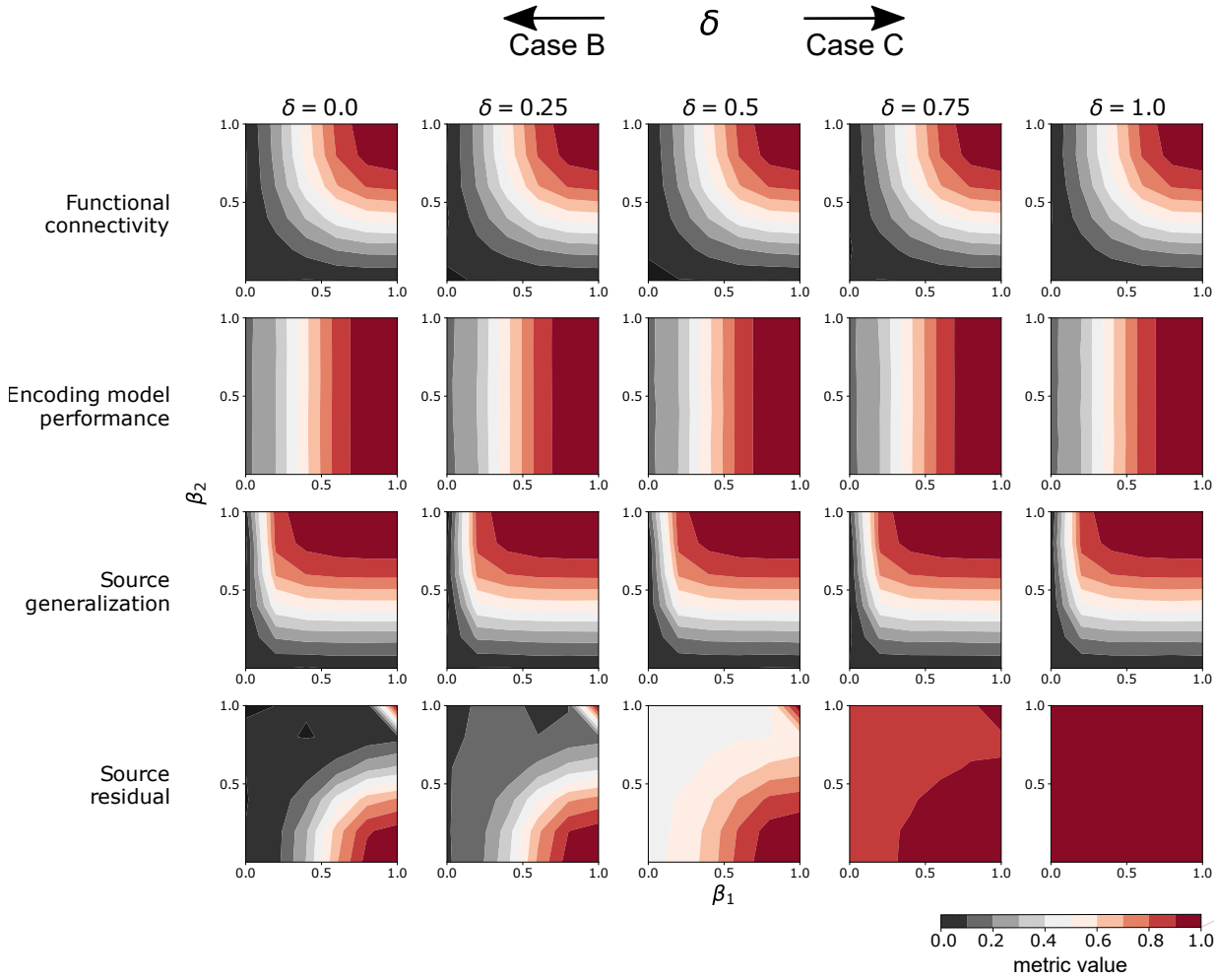

Figure 5: Extending on Fig. 3.2B, this figure shows how each metric varies under simulations performed at different signal-to-noise ratios as we vary δ, β_1, β_2 when $\alpha = 1.0$ is fixed.

A.5.2 Courtois NeuroMod

Results included in this manuscript come from preprocessing performed using fMRIPrep 20.1.0 [13, 14]. Three participants are native French speakers and three are native English speakers. All participants are fluent in English and report regularly watching movies in English.

A.5.3 Other Pre-processing

The fMRI datasets and Shen atlas were provided in different template spaces and voxel sizes. We resample and register the Shen atlas (MNI27 template space, voxel size = 1 mm isotropic) to both the HCP template space (MNI152NLin6Asym, voxel size = 1.6 mm isotropic) and the Courtois NeuroMod template space (ICBM2009cNlinAsym, voxel size = 2 mm isotropic) using FSL FMRIB Linear Image Registration Tool (FLIRT) [38]. We perform all analyses for the two datasets in their respective template space.

We further process the ELMo embeddings before we use them as the input features to our encoding models. First, we use a Lanczos filter with the same parameters as Huth et al. to downsample the embeddings into a feature matrix where each row corresponds to a feature vector for a TR [35]. Then, to reduce the dimensionality of our feature space we use principle component analysis (PCA) to select the first 10 principle components. The first 10 principle components explain 50.5% of the variance in the Courtois NeuroMod dataset and 49.9% of the variance in the HCP dataset. Next, to account for the lag in the hemodynamic response in fMRI data, we delay the feature matrix in accordance with previous work [35, 47, 65].

A.6 Additional Individual-Level Empirical Results Using Our Proposed Framework

We present the individual-level results for the remaining four participants in the Courtois NeuroMod dataset and four additional participants for the HCP dataset. We observe that these additional participants appear similar to the average and two representative individual participants presented in the main text.

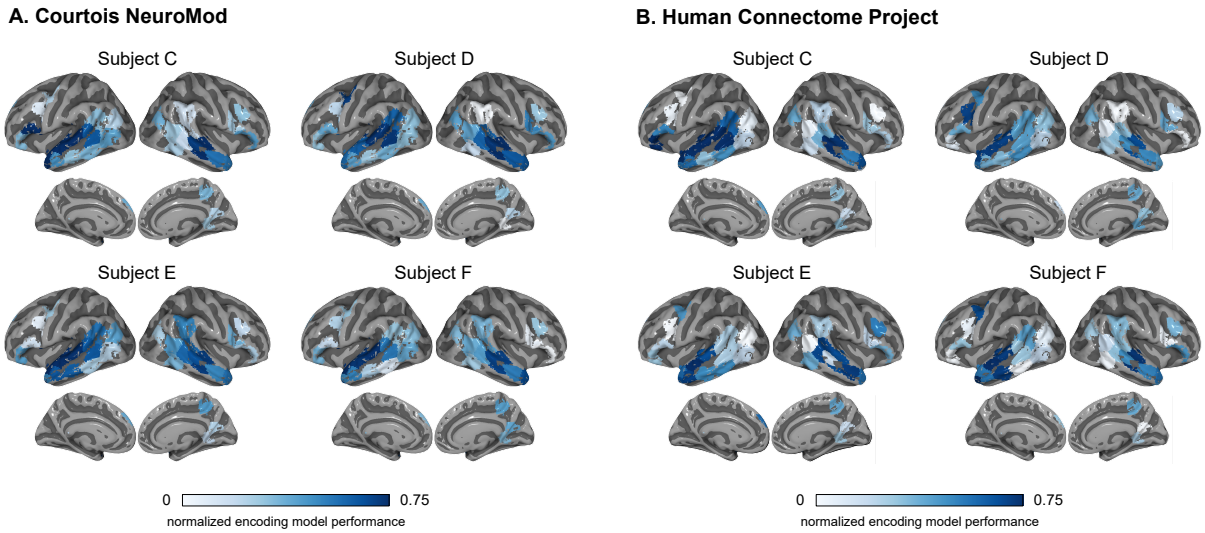

Figure 6: (Related to Fig. 3.3) Encoding Model Performance. Similar to Fig. 3.3 in the main text, this figure shows the normalized encoding model performance at 33 significantly predicted ROIs (corrected at level 0.05) for participants C-F in both the Courtois NeuroMod and Human Connectome Project datasets. Plots were created using the Pycortex software [21].

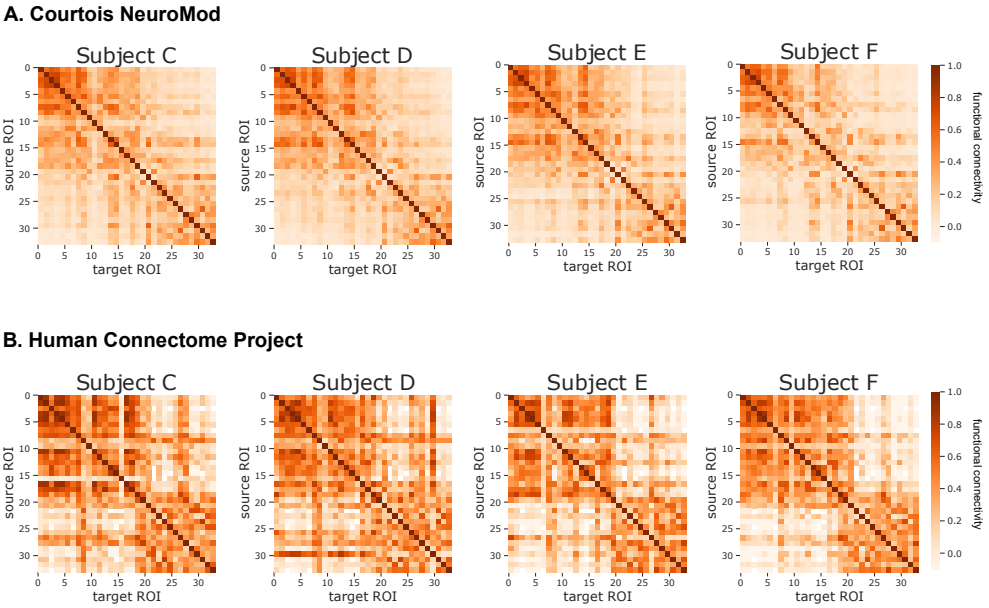

Figure 7: (Related to Fig. 1) Functional Connectivity. Similar to Fig. 1 in the appendix, this figure shows the significant pairwise correlations of the 33 language ROIs (corrected at level 0.05) for participants C-F in both the Courtois NeuroMod and Human Connectome Project datasets. The overwhelming majority of ROI pairs have significant correlations. This is consistent with the group level and individual participants presented in Fig. 1.

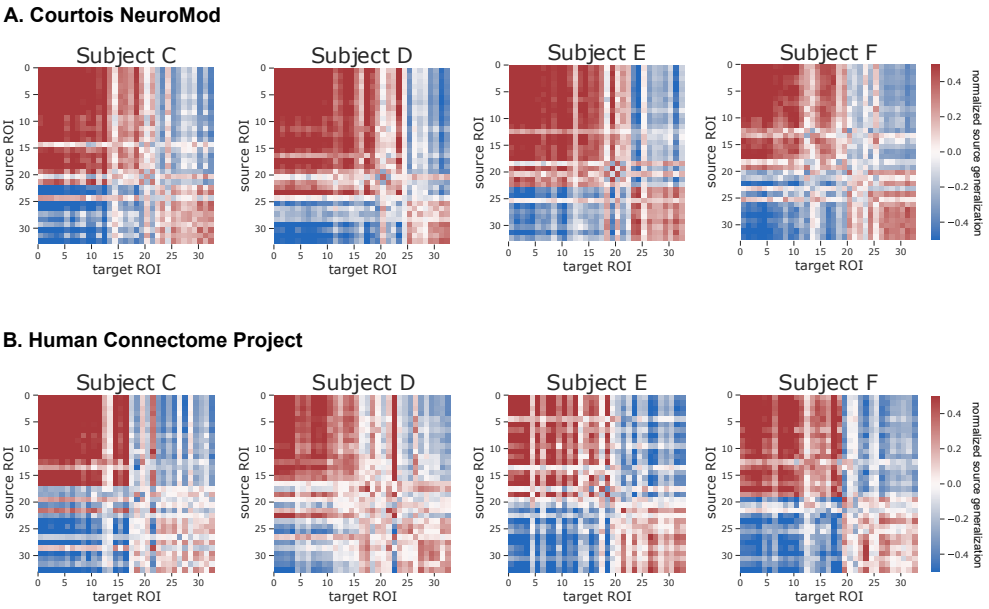

Figure 8: (Related to Fig. 3.4) Source Generalization. Similar to Fig. 3.4 in the main text, this figure shows the normalized source generalization for participants C-F in both the Courtois NeuroMod and Human Connectome Project datasets. ROI pairs with high normalized source generalization (red) are consistent across participants C-F in both datasets. They are also consistent with the group level and individual participants presented in the main text.

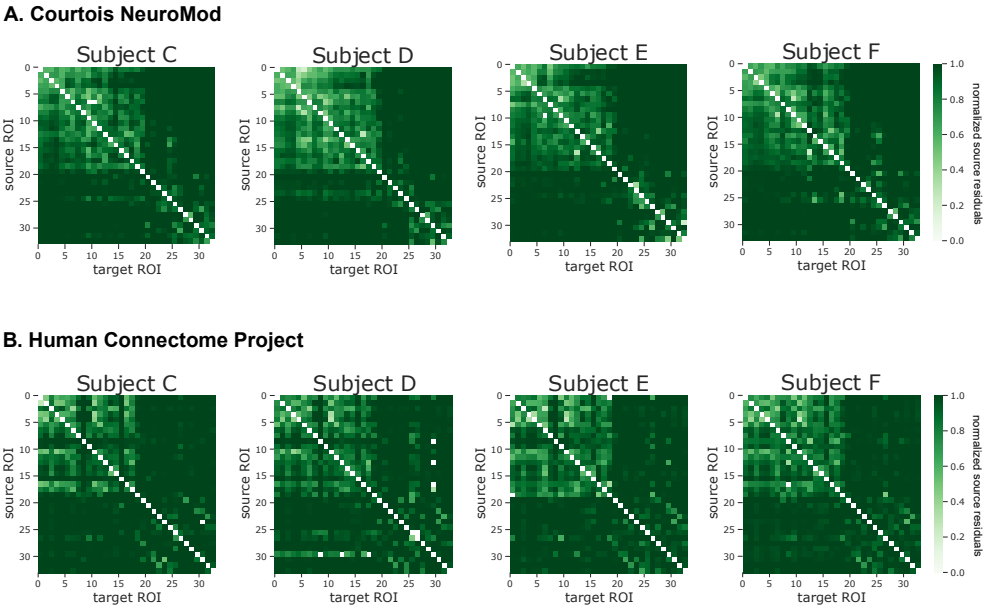

Figure 9: (Related to Fig. 3.5) Source Residuals. Similar to Fig. 3.5 in the main text, this figure shows the normalized source residuals for participants C-F in both the Courtois NeuroMod and Human Connectome Project datasets. ROI pairs with high normalized source residuals (green) are consistent across participants C-F in both datasets. They are also consistent with the group level and individual participants presented in the main text.

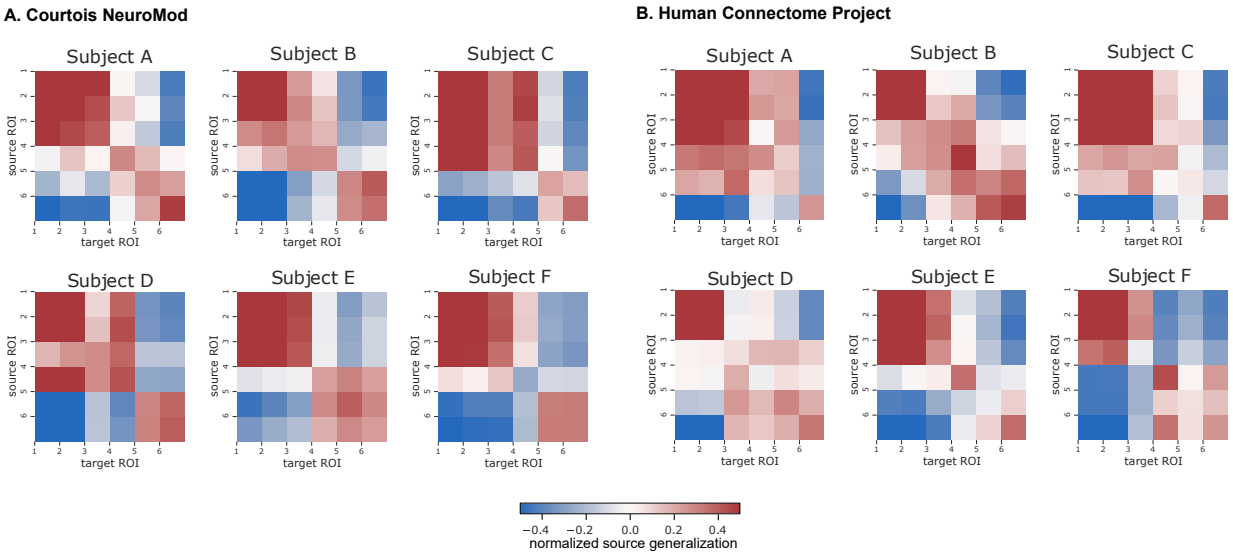

Figure 10: (Related to Fig. 3.6) Proposed Framework Example Individual Level Source Generalization. Similar to Fig. 3.6 in the main text, this figure shows the normalized source generalization for the six ROIs in the example using the proposed framework for participants A-F in both datasets. The ROI pairs with high normalized source generalization (red) are consistent across participants A-F in both datasets. They are also consistent with the group level presented in the main text.

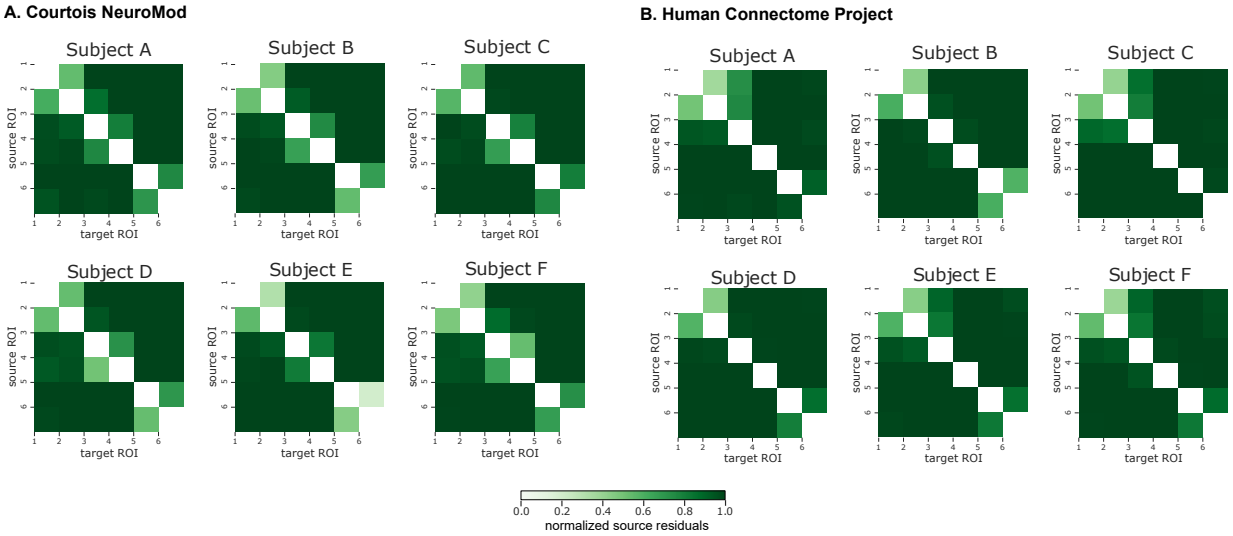

Figure 11: (Related to Fig. 3.6) Proposed Framework Example Individual Level Source Residuals. Similar to Fig. 3.6 in the main text, this figure shows the normalized source residuals for the six ROIs in the example using the proposed framework for participants A-F in both datasets. The ROI pairs with high normalized source residuals (green) are consistent across participants A-F in both datasets. They are also consistent with the group level presented in the main text.

A.7 Additional Individual-Level Comparisons between *Linear-Analytical*, *Linear-GD*, *MLP-GD* and the Noise Ceiling

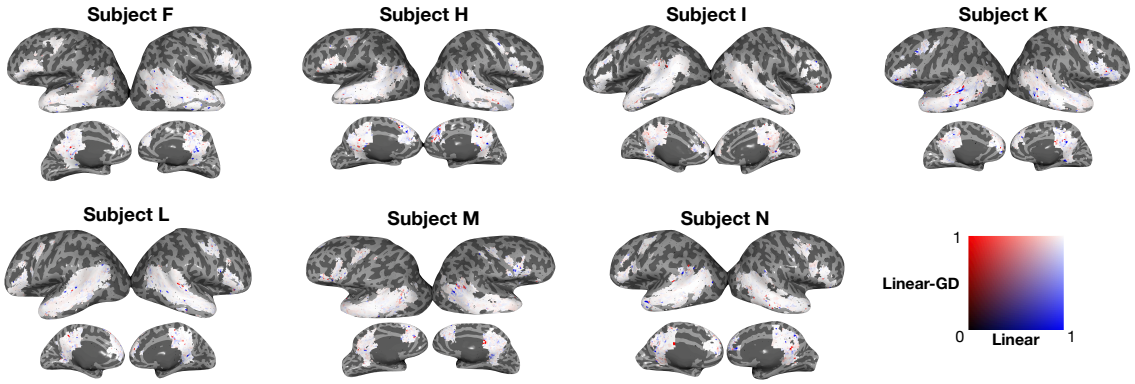

Figure 12: Extending on Fig. 4.2, which visualizes the encoding performance comparison between *Linear-Analytical* and *Linear-GD* for a sample subject (subject J), this figure shows the same comparison on each of the remaining participants from this study.

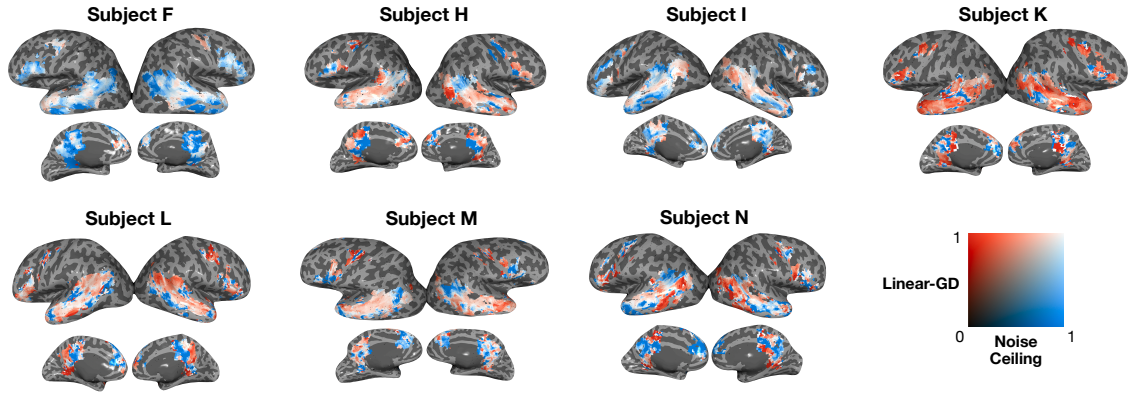

Figure 13: Extending on Fig. 4.3 and Fig. 4.5, which visualize the comparison of *Linear-GD*'s encoding performance with an estimated noise ceiling for a sample subject (subject J), this figure shows the same comparison for the remaining participants from this study.

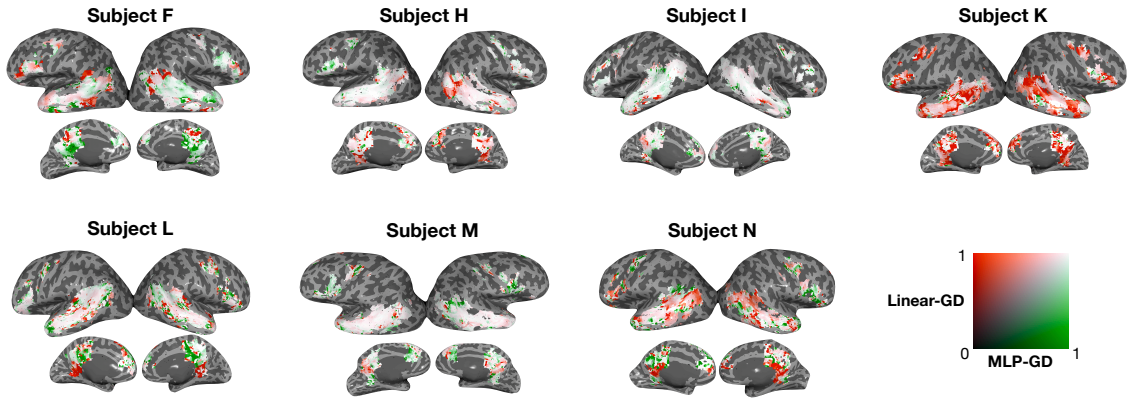

Figure 14: Extending on Fig. 4.4, which visualizes the encoding performance comparison between *Linear-GD* and *MLP-GD* for a sample subject (subject J), this figure shows the same comparison on each of the remaining participants from this study.