2020

Investigating the Source and Structure of Unexplained Variance in Natural Scenes fMRI Data

Maggie Mae Mell
Medical University of South Carolina

Investigating the Source and Structure of Unexplained Variance in Natural Scenes fMRI Data

by

Maggie Mae Mell

A dissertation submitted to the faculty of the Medical University of South Carolina in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the College of Graduate Studies.

Department of Neurosciences

2020

Approved by:

Thomas Naselaris
Chairman Advisory Committee

Jane Joseph

Andreana Benitez

Lisa McTeague

Truman Brown

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

A critical measure of understanding in neuroscience is the ability to predict how the brain will respond to arbitrary, complex stimuli. The visual cortex is a key target for elucidating the computations performed on incoming sensory information. Primates are highly dependent upon the visual system in generating perceptual experience, with a large portion of cortical area dedicated to visual processing. Additionally, decades of previous research have built a foundational body of knowledge on the structure and function of key visual areas, much more so than any other cortical sensory system (For review see Trenholm & Krishnaswamy, 2020).

In humans, neural activity in the visual system is often characterized using blood oxygenation level dependent (BOLD) functional magnetic resonance imaging (fMRI)(Logothetis, 2007, 2008). BOLD fMRI is a noninvasive, indirect measure of neural activity. Computational models that predict BOLD activity in response to visual stimuli are known as encoding models(David & Gallant, 2005; Dumoulin & Wandell, 2008; Kay et al., 2008; St-Yves & Naselaris, 2018; Wu et al., 2006). Typically, encoding models extract physical features of the visual stimulus to map onto BOLD activity recorded from individual voxels. Features may be simple, as in the presence and location of high contrast edges, or they may be far more complex such as the nonlinear transformations embedded in layers of an artificial neural network.

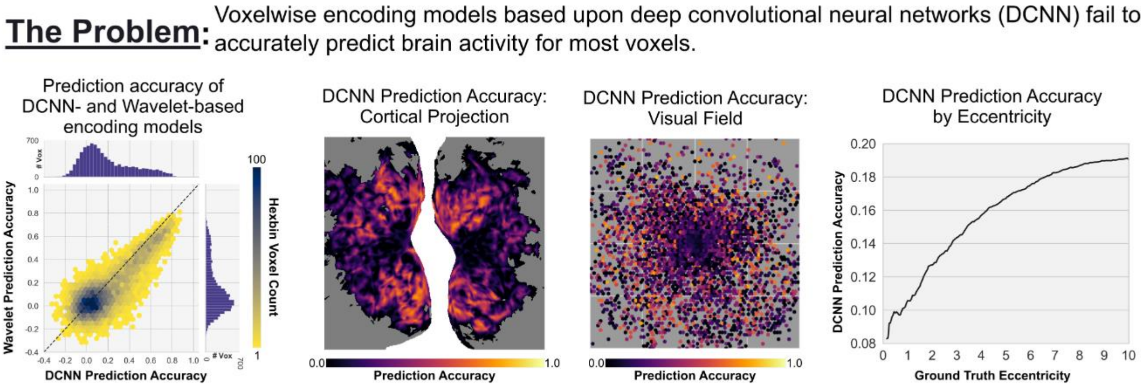

Figure 1.1 *Visualizations of the prediction accuracy of the DCNN-based encoding model.* **Left:** The joint distribution of prediction accuracy (Pearson correlation between predicted and measured brain activity) for the DCNN-based encoding model (x-axis) and Gabor wavelet-based encoding model (y-axis; data taken directly from St. Yves 2018). The slightly higher count (color, yellow=low count, dark blue = high count, white = no data) of voxels below the line at unity (dashed) reveals the advantage of the DCNN- over the wavelet-based encoding model. **Left Middle**: Prediction accuracy of the DCNN-based encoding model (color) projected onto a cortical flatmap. Prediction accuracy is poorest (dark purple) in the foveal representation. **Right Middle**: Prediction accuracy of the DCNN-based encoding model (color indicates median) projected into visual space (gray square) using the receptive locations (hexagonal bins) of all voxels. Prediction accuracy is poorest for voxels with foveal receptive fields (bins near center of square). **Right**: Cumulative average prediction accuracy of the DCNN-based encoding model (y-axis indicates median) against receptive field eccentricity (x-axis).

Voxel-wise encoding models are a flexible and powerful tool that allow researchers to replicate many findings from animal physiological literature in human fMRI paradigms(Cheng, 2018; Gaglianese et al., 2017; Grinvald et al., 2000). Currently the best encoding models for experiments utilizing natural scene stimuli are based on deep convolutional neural networks (DCNN) that have been trained on object recognition tasks(Güçlü & van Gerven, 2015; Kriegeskorte, 2015; St-Yves & Naselaris, 2018; Yamins et al., 2014). DCNN-based encoding models are impressive in their ability to accurately predict BOLD activity measured in voxels across the visual hierarchy, using a single underlying model of computation. Nonetheless, DCNN-based encoding models (and, to our knowledge, all encoding models) fail to accurately predict brain activity in response to natural scenes in most voxels in all visual areas (**Figure 1.1**).

This leads to an important unanswered question in the field of visual neuroimaging, and neuroimaging in general: Is there structure and meaning in the variance unexplained by the current best encoding models, or is this variance due to measurement-based noise and artifacts systemic to fMRI experiments? The answer to this question has repercussions both in how we generate computational models of the visual system as well as how we process and interpret fMRI data in general.

In theory, noise is unwanted, often random, information in a signal. In practice, noise is generally anything that cannot be directly attributable to the stimulus, whether it be physiological processes, scanner artifacts, or activity intrinsic to the brain. However neurophysiological research in animals challenges this line of thinking, showing that intrinsic or spontaneous neural activity is not just noise, but rather a potent and functionally relevant signal.

In macaque V1, spontaneous activity recorded with multi-electrode arrays showed a similar topological structure to activity recorded during natural scenes (Singh et al., 2008). Neural population activity in visual cortex of cats revealed spontaneous activity has similar amplitudes and spatial patterns as responses evoked by visual stimuli(Arieli et al., 1995; Tsodyks et al., 1999). Unexplained variance in natural scenes fMRI may be related to the intrinsic activity of the brain with structure and functional importance. Intrinsic activity consumes up to 90% of the brain's metabolic resources, therefore it would be highly inefficient if this activity did not serve a functional purpose(Raichle & Mintun, 2006).

Another indication that unexplained variance is not noise comes from resting-state fMRI (rs-fMRI) literature. Resting state fMRI refers to the spontaneous activity recorded while a subject is in the scanner with no external stimulus present(Biswal et al., 1995). Originally, rs-fMRI itself was thought to be uninteresting noise until the seminal 1995 study by Biswal and colleagues revealed that the intrinsic BOLD fluctuations recorded during rest are correlated between anatomically connected but distant areas. In the visual system, rs-fMRI functional connectivity networks have been found to be highly reproducible across studies and consistent with underlying anatomical structure(Friston, 2011; Van Den Heuvel & Pol, 2010). Importantly, resting state fluctuations reveal topographically mapped patterns of connectivity and show higher correlations with task-based functional connectivity patterns elicited by naturalistic stimuli as opposed to synthetic stimuli(Heinzle et al., 2011; Strappini et al., 2019).

The goal of this thesis is to investigate structure of unexplained variance of fMRI natural scene data with the hypothesis that endogenous activity is the main source of variance. We will use an existing dataset to explore how unexplained variance is shared between 6 visual areas, V1, V2, V3, V4, V3ab, LO. In Chapter 2 we describe our experimental methods in detail. We introduce the voxel-to-voxel modeling method, a unique approach that leverages the coactivations between visual areas to predict brain activity.

In Chapter 3 we address our first aim of building voxel-to-voxel encoding models of brain activity during natural scene viewing. We found voxel-to-voxel encoding models can successfully predict brain activity of target voxels for single-trial and repeated-trial data. Our second aim is addressed in Chapter 4 where we further investigate the source and structure of unexplained variance. We show the source of unexplained variance is shared across voxels but specific to each individual brain. Further, we find evidence of retinotopic structure in unexplained variance, even after removing stimulus-based signal.

In Chapter 5 we address our final aim of building pixel-to-pixel encoding models of unit activations from deep neural networks trained for object recognition to compare to results from Aims 1 and 2. We find the pattern of predictive connectivity is different from that seen in the hierarchy of brain areas. Chapter 6 presents preliminary results wherein we extend our modeling method to data acquired with 2-photon calcium imaging. Neuron-to-neuron models trained with data from mouse V1 share surprising similarities with lateral fMRI voxel-to-voxel models. Finally, Chapter 7 we summarize our findings and discuss potential avenues for further investigation.

Chapter 2: Experimental Methods

2.1 Data

2.1.1 Natural Scenes

Two datasets utilizing natural scene images as stimuli were analyzed separately. The main analyses were applied to the publicly available Vim-1 dataset (Kay et al., 2011). One additional analysis was applied to the Natural Scenes Dataset, which will be made available for public use in the coming year. Both datasets are described below.

Vim-1

In this experiment, described in detail in (Kay et al., 2008), two healthy males passively viewed 1870 natural scene photographs. Stimuli were greyscale and 500 x 500 pixels in size, subtending 20° of visual angle in each direction. A 4x4 pixel white square in the center of the image served as the fixation point. On each trial one photograph was presented for a total of 1 second. Prior to scanning stimuli were split into a training set of 1750 and a test set of 120. Each training image was presented twice and test images were presented 13 times within one scanning session. Each scan session consisted of 7 runs. There were 5 scan sessions in total. MRI data was acquired at University of California, Berkeley on a 4-Tesla scanner. The acquisition parameters are as follows: 18 2mm x 2.5mm x 2mm coronal slices, FOV 128mm x 128mm covering the occipital cortex, T2*-weighted gradient-echo EPI sequence, TR =1s, TE = 28ms.

Vim-1 data is publicly available as minimally pre-processed 4D Nifti images. To obtain activation amplitudes for each image in all voxels, we used the Rank-1 GLM with 3HRF basis estimation procedure detailed in (Pedregosa et al., 2015). Briefly, runs were separated by session, 5 sessions each containing 5 training runs and 2 testing. Local detrending using Savitzky–Golay filter (Savitzky & Golay, 1964) with a polynomial of degree 4 and window length of 671 seconds was applied to each run separately. HRF function and activation amplitude estimates were calculated for training runs in each session for every voxel. HRF functions were estimated using only training data and then applied to testing runs from the same session to obtain activation amplitude for test images. The mean and standard deviation for training runs were used to z-score amplitude estimates for both training and testing runs. Voxels were localized to regions of interest (ROI) including V1, V2, V3, V4, V3a, V3b, and LO (for our analyses we combined V3a and V3b into one area, V3ab).

Natural Scenes Dataset

The Natural Scenes Dataset was acquired on a 7T Siemens Magnetom scanner at Center for Magnetic Resonance Research, University of Minnesota. Data were collected with gradient-echo EPI, 1.8mm isotropic voxels across the entire brain. Data from one subject in this unpublished dataset was used for one preliminary analysis. In this experiment subjects viewed stimuli drawn from the publicly available COCO dataset. Each subject saw 10,000 distinct images, each image was presented 3 times over 40 scan sessions for a total of 30,000 trials. In additional to the functional task data, 20 resting state runs were acquired using the same parameters. Preprocessed resting state time series and beta activation amplitudes for each of the 30,000 trials were provided to us by a collaborator. Full details regarding acquisition, preprocessing and amplitude estimation will be available in the forthcoming paper. More information can be found here: http://naturalscenesdataset.org/

2.1.2 Retinotopic Mapping

One subject that participated in the original vim-1 dataset collection (S1) was also scanned to collect retinotopic data for ground truth receptive field estimation. Retinotopy data was collected using standard rotating wedge, expanding ring, and drifting bar stimuli (Dumoulin & Wandell, 2008; Kay et al., 2013). Retinotopic stimuli are 768 x 768 pixels and subtend 22.1 degrees in each direction. 14 separate runs of data were collected, 7 rotating wedge/expanding ring and 7 drifting bar.

Retinotopy data was acquired on a 7T scanner at The University of Minnesota's Center for Magnetic Resonance Research. Functional data consisted of 66 coronal slices with .8 mm isotropic voxels, FOV 160mm x 160mm, covering the posterior half of the brain. A T2*-weighted, multiband slice interleaved, gradient-echo EPI sequence was used with the following parameters: TR = 1.5s, TE= 22.2ms, multiband acceleration factor = 3.

The retinotopy data was preprocessed using FSL's FEAT (Jenkinson et al., 2012; Woolrich et al., 2001). Additionally, FLIRT and FNIRT were used to bring all 16 scans into alignment via linear rigid-body transformations and non-linear warping(Andersson et al., 2010; Jenkinson & Smith, 2001). For direct comparisons to the vim-1 data set, the 7T data was down-sampled to 1.5mm isotropic voxels and aligned into vim-1 space. All alignments across preprocessing and down-sampling strategies were concatenated into one interpolation step and applied directly to the filtered functional data resulting from FEAT. Receptive field location for each voxel was estimated using population receptive field analysis (AnalyzePRF), procedure detailed in (Dumoulin & Wandell, 2008; Kay et al., 2013)

Surface reconstruction

A structural T1 volume acquired during the 7T retinotopy experiment was skull-stripped and passed to Freesurfer's recon-all (version 6) (Dale et al., 1999) for surface reconstruction procedures. Relaxation cuts on the inflated cortical surface were made in Blender (v2.78) (Community, 2018) and then imported back into Freesurfer for flattening. Finally, all surfaces were imported into pycortex for rendering of cortical flatmaps. Results in functional data format were rigidly aligned to the structural T1 with FSL FLIRT (Jenkinson & Smith, 2001; Jenkinson et al., 2012) and projected onto cortical surfaces in pycortex (Gao et al., 2015).

2.1.3 2-Photon Calcium Imaging

An additional preliminary analysis was applied to publicly available data recorded from mouse V1 using 2-photon calcium imaging (Stringer et al., 2019a). The dataset included fluorescence response values to natural scene stimuli recorded from ~10000 neurons in 8 mice. Full details regarding the experiment and acquisition methods can be found in (Stringer et al., 2019b). Data from one mouse was used for this analysis.

2.2 Encoding Models

2.2.1 Stimulus-to-voxel fWRF

For all voxels in the vim-1 natural scenes dataset, the feature-weighted receptive field model (fwRF) was applied to the feature maps of a deep convolutional neural network (**Figure 2.1** top). Full details regarding this model can be found in (St-Yves & Naselaris, 2018). Briefly, the fwRF is a form of voxel-wise encoding model that separates the specification and estimation of receptive field location and size from feature tuning. The fwRF uses the following model to generate predictions of brain activity, $\hat{r}_t$, in response to a visual stimulus :

$$\hat{r}_t = \sum_{k=1}^{K} w_k \int_{-D/2}^{D/2} \int_{-D/2}^{D/2} g(x, y; \mu_x, \mu_y, \sigma_g) \phi_{i(x)j(y)}^{k}(S_t) dxdy$$

Where D is the visual angle sustained by the image, the function $\phi_{i(x)j(y)}^{k}$ specifies the value of pixel (i, j) of the k^{th} feature map applied to the stimulus S_t, and $g(x, y; \mu_x, \mu_y, \sigma_g)$ is the feature pooling field, which is an isotropic-2D Gaussian function, with center (μ_x, μ_y) and radius σ_g. The feature pooling field indicates the region of visual space in which stimulus variation induces variations in activity of the voxel. The feature weights, w_k, indicate the features encoded in the activity of the voxel. The set of feature maps used are the same for each voxel, but the weights assigned to each feature will vary.

In this paper, the features for the stimulus-to-voxel model were the feature maps of AlexNet (Krizhevsky et al., 2012), a DCNN with one input layer, five convolutional layers and three fully-connected layers. AlexNet is trained to classify images in the ImageNet database, a pre-trained version was downloaded from the Caffe Model Zoo. The location and radius of the feature-pooling field, as well as the feature weights are estimated by minimizing the sum-of-squared prediction error between model output and brain activity for each voxel over the set of image/response pairs in the training set. Values for the location and radius of the feature-pooling field, i.e. the fwRF center and size, are inferred via a brute force search through a grid of candidate locations and radii. Values for the feature weights are estimated using ridge regression

2.2.2. Voxel-to-voxel

Voxel-to-voxel models (**Figure 2.1** middle) linearly combine activity from one brain area to predict activity in one voxel:

$$\hat{r}_t^{target} = W\boldsymbol{r}_t^{source}$$

where $\hat{r}_t^{target}$ is the predicted activation of the target voxel, W is a matrix of model weights, and $\boldsymbol{r}_t^{source}$ an array of activations from source voxels. We used ridge regression to determine the weights assigned to each voxel in a source area. We fit separate voxel-to-voxel models for each pair of visual areas named above. Thus, for each target voxel we fit six distinct voxel-to-voxel models corresponding to the six ROIs named above.

For each pair of ROIs we refer to a voxel-to-voxel model as “feedforward” if the source voxels are lower in the hierarchy of ROIs than the target voxel. We refer to a voxel-to-voxel model as “feedback” if the source voxels are higher in the hierarchy than the target voxel. We refer to a voxel-to-voxel model as “lateral” if the source and target voxels are in the same ROI . The hierarchy of ROIs is defined by the sequence V1, V2, V3, V4, LO/V3ab, where V1 is the “lowest” ROI in the hierarchy

2.2.3 Pixel-to-Pixel

Pixel-to-pixel models (**Figure 2.1** Bottom) linearly combine activity in the pixels of one DCNN layer to predict activity of a single target pixel:

$$\hat{\phi}_t^{target} = W\boldsymbol{\phi}_t^{source}$$

where $\hat{\phi}_t^{target}$ is the activation of a target pixel in one of the feature maps of the DCNN, $W$are the pixel-to-pixel model weights, and $\boldsymbol{\phi}_t^{source}$ is an array of activations of source pixels taken from all feature maps in one layer of the DCNN. The DCNN was built and trained in house. The network consists of 5 convolutional layers with a rectifier non-linearity and one fully connected layer. The network was trained to classify based on the 10 categories indicated in the CIFAR-10 dataset(Nishida et al., 2019). As with voxel-to-voxel models, we fit a pixel-to-pixel model for every possible pair of source layer and target layer, and refer to pixel-to-pixel models as feedforward, feedback and lateral depending upon the relative positions of the source and target layers in the network hierarchy. Note, due to computational constraints, lateral models were calculated slightly differently for pixel-to-pixel models. Specifically, 10% of pixels in a layer were randomly selected as target pixels, with the remaining 90% of pixels as the source pixels. This procedure was repeated ten times such that a lateral model was computed for every pixel in a layer. Due to the high redundancy of feature information in DCNN layers, we do not expect this procedure to affect the lateral model prediction accuracy. We chose to use this DCNN, rather than AlexNet used in the encoding model to reduce computational time and allow for complete pixel-to-pixel models of each layer. However, a similar analysis was performed on sub-samples of AlexNet layers.

2.2.4 Neuron-to-Neuron

In a preliminary analysis applied to a newly publicly available dataset, we extended our methodology to single neuron activation values. Neuron-to-neuro models linearly combine activity in one neuron of mouse V1 predict activity of a single target neuron:

$$\hat{\psi}_t^{target} = W\boldsymbol{\psi}_t^{source}$$

where $\hat{\psi}_t^{target}$ is the predicted activation of the target neuron, W is a matrix of model weights, and $\boldsymbol{\psi}_t^{source}$ an array of activations from all other neurons. This model is conceptually akin to the lateral modes in the voxel-to-voxel analysis.

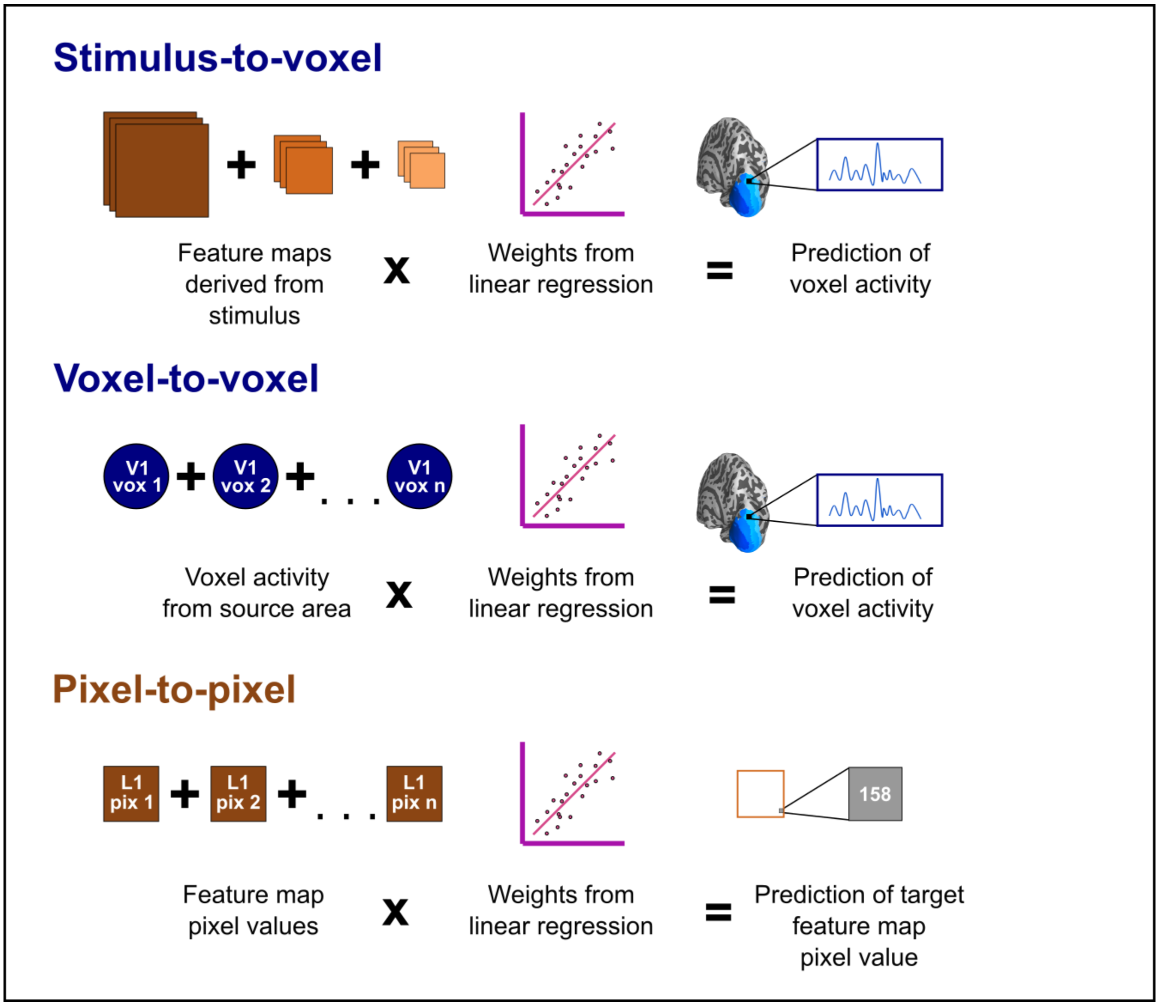
Stimulus-to-voxel
Feature maps derived from stimulus
x
Weights from linear regression
=
Prediction of voxel activity
Voxel-to-voxel
V1 vox 1
+
V1 vox 2
+
. . .
V1 vox n
Voxel activity from source area
x
Weights from linear regression
=
Prediction of voxel activity
Pixel-to-pixel
L1 pix 1
+
L1 pix 2
+
. . .
L1 pix n
158
Feature map pixel values
x
Weights from linear regression
=
Prediction of target feature map pixel value

Figure 2.1 *Model Types* **Top**: The DCNN-based encoding model is a stimulus-to-voxel model that transforms stimuli into a set of feature maps (brown squares) and then into a prediction of voxel activity (blue curve). In the DCNN-based encoding model the transformation of stimuli into feature maps is performed by a deep neural network; the transformation from feature maps to voxel activity is estimated via linear regression (idealized pink line). **Middle**: In a voxel-to-voxel model activity in a population of source voxels (blue circles) is linearly transformed into a prediction of activity in a target voxel. **Bottom**: In a pixel-to-pixel model activity in a population of source pixels in a feature map of the DCNN (brown squares) is linearly transformed into a prediction of activity of another target pixel in the DCNN.

2.2.5 Training, cross-validation, prediction accuracy

Voxel-to-voxel repeated-trial encoding models were trained on 1750 responses to natural scene photographs and tested on the remaining 120. Pixel-to-pixel models were trained on pixel activation values to the same stimuli in the vim-1 dataset. Neuron-to-neuron models were trained on a randomly selected 90% (N=2520) of images and testing on the remaing 10% N=(280).

Ridge regression hyper-parameter values were selected via line search by cross-validating against 20% of the training data. Prediction accuracy is the Pearson correlation between model predictions and measured activity (in the brain for voxel-to-voxel models, in the DCNN for pixel-to-pixel models).

2.2.6 Single-trial Analyses

Voxel-to-voxel, stimulus-to-voxel fwRF, and neuro-to-neuron models were also applied to single-trial activation values. In these analyses voxel encoding models were trained on 3500 responses (1750 stimuli x 2 trials) and tested on 1560 responses (120 stimuli x 13 trials). Neuron-to-neuron data consisted of 2 trials per stimulus, models were trained on 90% of the total trials (N=5040) and tested on the remaining 10%(N=560), ensuring trials from the same stimulus did not appear in both training and testing sets. A second voxel-to-voxel and neuron-to-neuron 'Mix' analysis was also applied. Mix models are trained to predict the opposite trial, i.e. source activations from trial one are trained to predict a target activation to trial 2 and vice versa. Each trial repetition in the stimuli test set is tested against every other repetition.

Chapter 3: Validation of voxel-to-voxel models

Specific Aim 1: *Build voxel-to-voxel encoding models of brain activity during natural scene viewing.* Hypothesis: Voxel-to-voxel encoding models can successfully predict brain activity of target voxels for single-trial and averaged trial data.

3.1 Overview and Rationale

To investigate the source and structure of unexplained variance in fMRI, we must apply a new type of encoding model. With the aim of capturing variance potentially unrelated to stimulus evoked activity, this model needs to be at least partially independent of stimulus presented. That is, specific visual characteristics associated with the stimuli should NOT be part of the model parameters. Additionally, to determine the scale and source of unexplained variance, this model should cover multiple hierarchical directions and spatial scales. As such, we will apply a voxel-to-voxel modelling method to an existing natural scene dataset.

If the source of unexplained variance is due to endogenous brain activity, we would expect voxel-to-voxel encoding models to leverage this activity to produce highly accurate predictions of stimulus response regardless of source-target pairing. Conversely, if unexplained variance was largely due to nuisance sources, we expect voxel-to-voxel models to produce accurate predictions and receptive field location only when source voxels were spatially adjacent to the target voxel. Therefore, the objective of this aim is to build a model that determines if the variance unexplained by stimulus-to-voxel models can be explained by the activity of other voxels and the extent across the visual hierarchy the source of unexplained variance is shared.

3.2 Methods

We used measurements of fMRI BOLD activity as participants passively viewed natural scene images. Each image in the training set was repeated twice, each image in the validation set was repeated 13 times. We analyzed image-specific activation values calculated from single-trial data, as well as data averaged over all trials for that image. We built voxel-to-voxel encoding models for every voxel (for specific details regarding voxel-to-voxel model building, please see Chapter 2: Experimental Methods). We fit separate voxel-to-voxel models for each pair of visual areas in the following ROIs: V1, V2, V3, V4, V3ab, & LO. For each pair of ROIs, we consider a model as 'feedforward' if the source voxels are lower in the hierarchy of ROIs than the target voxel; 'feedback' if the source voxels are higher than the target voxel; and 'lateral' if the source and target voxels are in the same ROI. The hierarchy is ordered as above, with 'V1' being the lowest ROI and 'V3ab/LO' both occupying the highest level. For comparison to stimulus-to-voxel models, we applied the feature weighted receptive field (fwRf) encoding model to each voxel.

All repeated-trial, activation values averaged over all trials, models were trained on 1750 natural scene responses in the training dataset and tested on a held-out set of 120 validation images. All single-trial models were trained on 3500 responses (1750 images x 2 repetitions) and tested on 1560 responses (120 images x 13 repetitions). Prediction accuracy was calculated as the Pearson correlation coefficient (CC) between model predictions and measured activity. Median prediction accuracy values for each target area were compared to stimulus-to-voxel control analyses and deemed successful if they are at least as good as stimulus-to-voxel models. See Chapter 3 on Experimental Methods for more details regarding experimental design, data acquisition, processing, and encoding models.

3.3 Results

3.3.1 Comparison of voxel-to-voxel encoding model to stimulus-to-voxel encoding model

To investigate unexplained variance, we must first ensure voxel-to-voxel models are able to account for significantly more variance than stimulus-to-voxel encoding models. Currently, the most accurate stimulus-to-voxel encoding models for predicting brain activity to natural scenes are based upon deep convolutional neural networks that have been trained on object recognition tasks. Our lab has developed one such model, the feature weighted Receptive field model (fwRF), we use this as our reference model. We use prediction accuracy, defined as the Pearson correlation coefficient between predicted and measured activity, as the comparison metric. In **Figure 3.1** we compare median prediction accuracy of each source-target voxel-to-voxel model (x-axis) to the median stimulus-to-voxel fwRF prediction accuracy in the target area (y-axis). Both repeated-trial and single-trial analyses are plotted for each subject. The coral line is the line of equality. Every voxel-to-voxel model falls on the right hand side of this line, indicating that the median amount of variance explained for any visual ROI is larger in a voxel-to-voxel model than in the stimulus-to-voxel fwRF model.

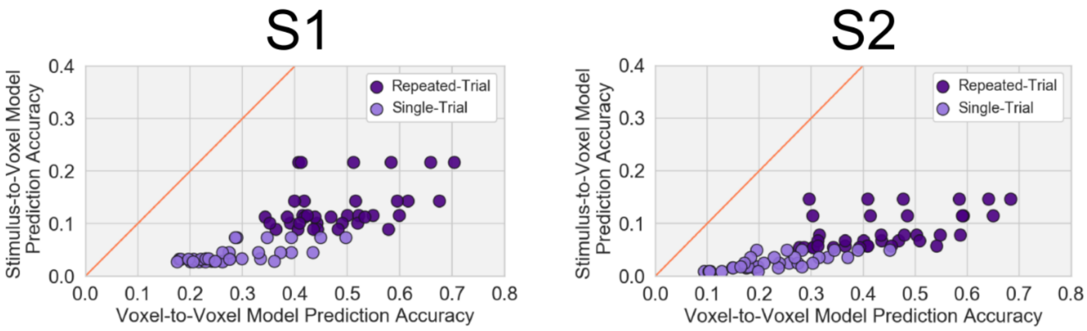

Figure 3.1 *Voxel-to-voxel models out predict stimulus-to-voxel models.* Each dot represents one source-target pairing. Median voxel-to-voxel prediction accuracy for target area plotted on each x-axis. Target area median stimulus voxel-to-voxel prediction accuracy on y-axis. Repeated-trial results appear in dark purple, single trial results are in light purple. S1 on left; S2 on right. For every source-target pairing in both subjects and trial-types, the voxel-to-voxel model out predicts the stimulus-to-voxel model for the target area.

Next, we recorded the number of voxels passing a prediction accuracy threshold for each model (repeated-trial 0.2, single-trial 0.08, both $p < 0.01$). To obtain the threshold we found the null distribution of prediction accuracy via permutation testing and selected the value three standard deviations from the mean of the distribution. Permutation testing was performed by shuffling a model's predicted activity over stimuli for each voxel and finding the correlation coefficient with the measured activity for that voxel, this assumes no relationship between model predictions and measured data. This process was repeated 10,000 times for every voxel in each model to build the null distribution. In both repeated-trial and single-trial analyses voxel-to-voxel models produce many more voxels passing threshold than the stimulus-to-voxel fwRF model (**Table 3.1**).

Table 3.1

	Voxels Passing Threshold Total N = 8470			
Model Source	**Repeated-Trial N**	**Repeated-Trial %**	**Single-Trial N**	**Single-Trial %**
Stim-fwRF	3261	39%	2882	34%
V1	7269	86%	7461	88%
V2	7427	88%	7485	88%
V3	7536	89%	7548	89%
V4	7477	88%	7565	89%
V3ab	6922	82%	7251	86%
LO	7041	83%	7343	87%

Finally, we assessed the pattern of prediction accuracy across the cortical surface. Stimulus-to-voxel models, including the fwRF, often fail in voxels located in areas processing foveal representations. **Figure 3.2** (top) shows each source area voxel-to-voxel model prediction accuracy for all voxels mapped onto the flattened cortical surface of Subject 1. Repeated-trial voxel-to-voxel models have high prediction accuracy across the cortical surface. To further this point, **Figure 3.3** plots prediction accuracy of lateral repeated-trial voxel-to-voxel models as a function of receptive field location. On the left prediction accuracy for each voxel is plotted according to receptive field location in visual space, on the right prediction accuracy is plotted as a function of eccentricity (the distance from the fovea to the receptive field). Both plots indicate high prediction accuracy across all receptive field locations. **Figure 3.2** (Bottom) shows single-trial analyses have relatively high prediction accuracy across the cortical surface, but do not reach the same level of accuracy as repeated-trial models, possibly due to the inherent variability of single-trial data.

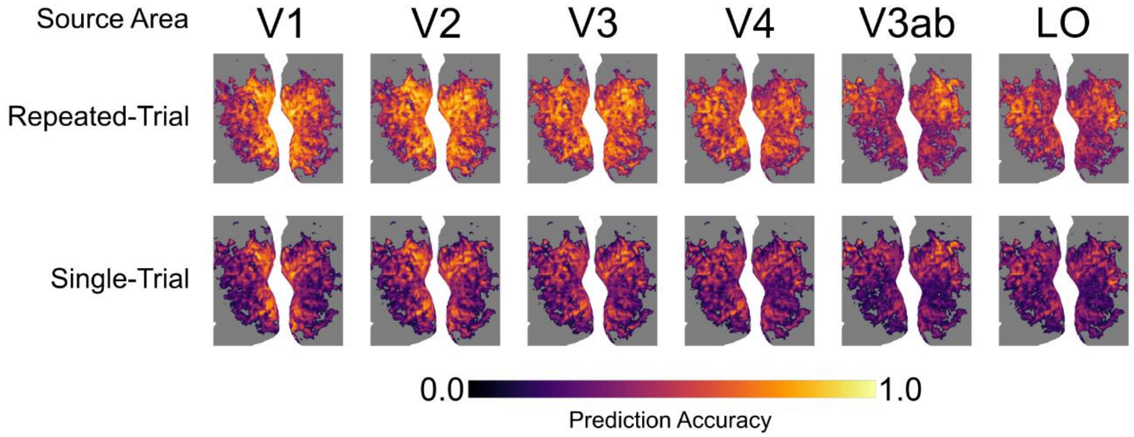

Figure 3.2 *Voxel-to-voxel models predict activity with high accuracy across visual cortex.* Cortical flatmaps of visual cortex shown for S1. Source areas indicated in top row, flatmaps below heading represents prediction accuracy projected onto the cortical flatpmap in all models for that source area. Repeated-trial analyses are in the top row, single-trial analyses in the bottom row. The color-scale starts at 0.0 with dark black and ends at 1.0 with bright yellow.

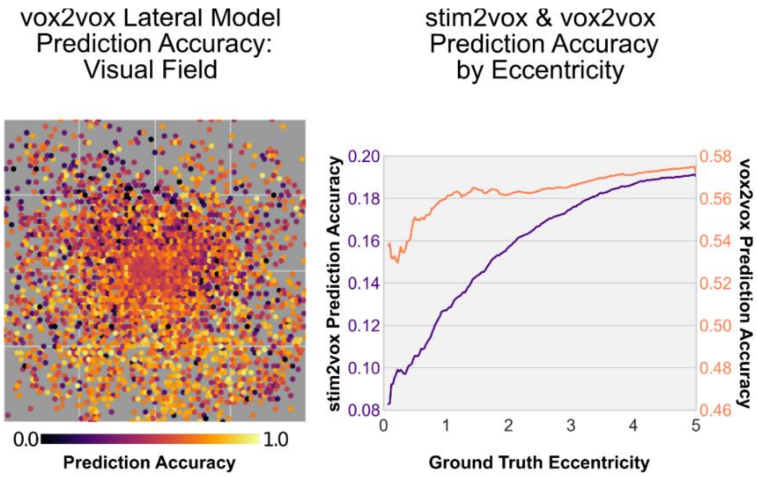

Figure 3.3 *Voxel-to-voxel models predict activity with high accuracy across visual field and eccentricities.* **Left:** Prediction accuracy of the lateral vox2vox models projected into visual space (format as in Figure 1.1). Prediction accuracy is roughly uniform across visual space. **Right:** Prediction accuracy against receptive field eccentricity (format as in Figure 1.1). Prediction accuracy for the voxel-to-voxel model(orange, right y-axis) is more uniform across eccentricities and generally higher than predication accuracy for the stimulue-to-voxel model (purple, left y-axis)

3.4 Conclusion

Both repeated-trial and single-trial voxel-to-voxel models account for significantly more variance than their stimulus-to-voxel fwRf counterparts. Repeated-trial models show high prediction accuracy across the entire cortical surface and across visual space. We conclude voxel-to-voxel models are a valid method for investigating unexplained variance in natural scenes data. In the following chapters we examine the parameters of each model to determine if there is structure and meaning in the additional explained variance.

Chapter 4: Unexplained variance is structured and shared across visual hierarchy

Specific Aim 2: *Investigate source and structure of unexplained variance in fMRI signal of brain activity.* Hypothesis: The source of unexplained variance is shared across voxels but specific to each individual brain. Within-brain prediction accuracy will be dependent upon the hierarchical signed distance between source and target voxels. Highly predictive source voxels will have overlapping receptive fields with that of target voxels.

4.1 Overview and Rationale

There are multiple questions that can be answered regarding the source and structure of variance unexplained by voxel-to-voxel models. Is there a unique source of unexplained variance specific to each voxel, or a common source shared across voxels? Are sources of unexplained variance specific to individual subjects or shared between them? Does any additional variance explained by voxel-to-voxel models adhere to known retinotopic principles? Is the predictive capabilities of voxel-to-voxel models dependent upon the source and target positions in the visual hierarchy? Our objective in Aim 2 is to utilize the parameters of voxel-to-voxel models to answer each of these questions.

If much of the variance unexplained by the stimulus-to-voxel model can be explained by the voxel-to-voxel model across all voxels and model pairings, we can infer that the causes of unexplained variance affect both target and source. Endogenous sources of activity are unique to each individual brain, therefore voxel-to-voxel models should not be more predictive than stimulus-to-voxel models when the source and target voxels are extracted from different brains. While we expect within-subject voxel-to-voxel models to be successful under all conditions, endogenous sources of activity like feedback processes, would produce an asymmetrical pattern of prediction accuracy based upon source distance and direction.

Endogenous sources of activity are expected to follow known properties and structure of the visual system and we would therefore expect highly predictive source voxels to share the same receptive field information as the target voxel. Removing activity attributable to physiological and machine noise should not degrade the predictive capability of voxel-to-voxel models, while averaging out spatial structure should.

4.2 Methods

We used the voxel-to-voxel models fit in Chapter 3 to determine the extent to which variance explained by stimulus-to-voxel models can be explained by the activities of other voxels. First, we compared voxel-to-voxel prediction accuracy to stimulus-to-voxel prediction accuracy across all voxels to determine if the source of unexplained variance is shared. Next, we used source voxels in one individual's brain to predict activity in target voxels of another individual's brain to determine if the source of unexplained variance is endogenous. We then compared median target area prediction accuracy in relation to the signed hierarchical distance between source and target areas.

To investigate structure of unexplained variance we first extracted the weight parameters from voxel-to-voxel models to read out the target voxel's receptive field location using the receptive field locations estimated from the stimulus-to-voxel fwRF model. We compared the receptive field read out by the voxel-to-voxel models with the receptive fields estimated by the ground truth retinotopic mapping experiment.

Finally, we applied voxel-to-voxel modelling to single-trial and control analyses. Control analyses included selecting a limited random number of voxels in each source area, adjusting for physiological and machine noise signals, and investigating the effect of removing stimulus related signal.

4.3 Results

4.3.1 Unexplained variance is shared within and between visual areas but is subject-specific.

For each source-target pairing, we visualized the joint distribution of stimulus-to-voxel fwRF and voxel-to-voxel model prediction accuracy. Even though voxel-to-voxel models are linear, voxel-to-voxel models have higher cross-validated prediction accuracy than the stimulus-to-voxel fwRF model for nearly every target voxel in every source/target pairing **Figure 5.1 (Left)**. These results show that, for example, the activity in V4 under an optimized linear transformation more

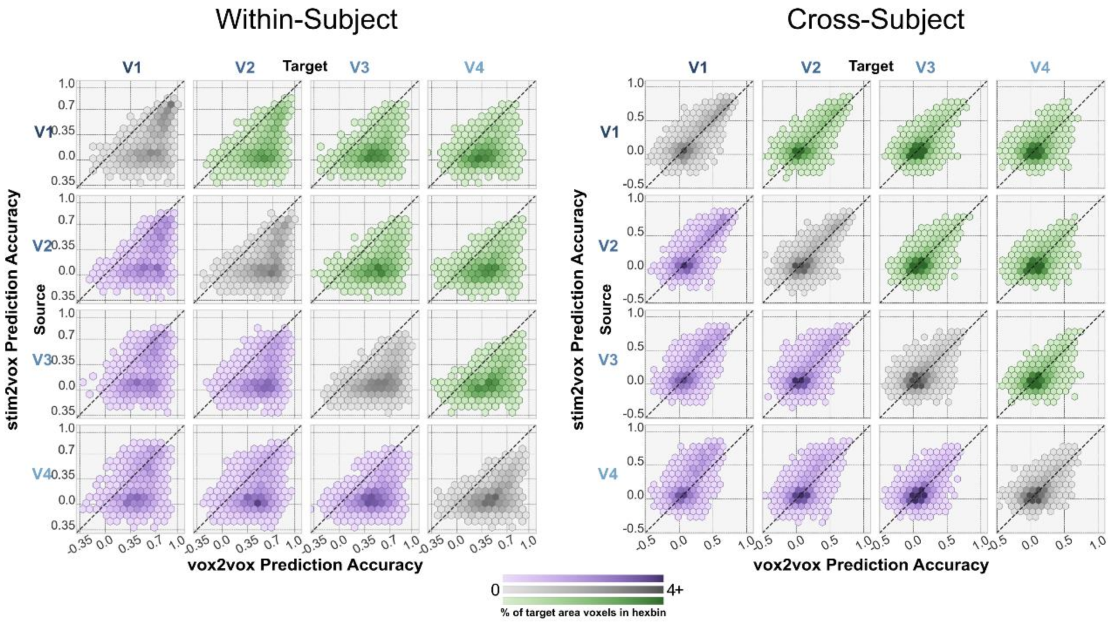

Figure 4.1 *Visualizations of the prediction accuracy of voxel-to-voxel models.* **Left:** Comparison of prediction accuracy of the stimulus-to-voxel (y-axis of each panel) and voxel-to-voxel models (x-axis) across a matrix of source (rows) and target (columns) visual areas (V1, V2, V3, V4). The voxel-to-voxel model generates more accurate predictions than the stimulus-to-voxel model for most voxels (percentage of voxels in source area indicated by color intensity, light = low, dark = high) for all source/target pairs and all voxel-to-voxel model types (green = feed-forward model, gray = lateral models, purple = feedback models). Data shown represents results from both S1 and S2. **Right:** Cross-subject voxel-to-voxel models do not enjoy the relative increase in prediction accuracy over stimulus-to-voxel fwRF models as same-subject voxel-to-voxel models. Data shown represents both cross-subject directions (S1 predicting S2 and S2 predicting S1)

accurately predicts activity in V1 than the stimulus under an optimized nonlinear transformation. Thus, the source of the variance that the stimulus-to-voxel fwRF model does not explain is clearly common to many voxels**.**

Next, we fit linear voxel-to-voxel models for source and target voxels in different brains. These cross-subject voxel-to-voxel models did not enjoy the dramatic improvement in prediction accuracy over the stimulus-to-voxel fwRF encoding model that we observed when within-subject voxel-to-voxel models were applied **Figure 4.1 (Right)**. This indicates that the cross-subject voxel-to-voxel models are, like stimulus-to-voxel models, blind to a source of variance that is common to voxels in the same brain.

4.3.2 Unexplained variance is retinotopically mapped

Is the source of variance unexplained in any one voxel common to *all* voxels in the same brain, or only to voxels that have overlapping receptive fields? To answer this question, we determined if voxel-to-voxel models preferentially connect target voxels to source voxels with receptive fields that overlap the target voxels'. To make this determination we plotted the weights of individual voxel-to-voxel models according to their receptive field locations, as estimated by the stimulus-to-voxel fwRF model. Importantly, we restricted our analysis to target voxels for which the prediction accuracies of the stimulus-to-voxel fwRF and voxel-to-voxel models were below and above a common threshold, respectively (Pearson correlation = 0.2; $p < 0.01$, permutation test). In other words, we analyzed only voxels that were “rescued" from the ball of nothingness by their respective voxel-to-voxel models. We found that the source voxels with the largest positive voxel-to-voxel model weights had receptive field locations that tended to cluster near the receptive field location of the target voxels (**Figure 4.2** Top).

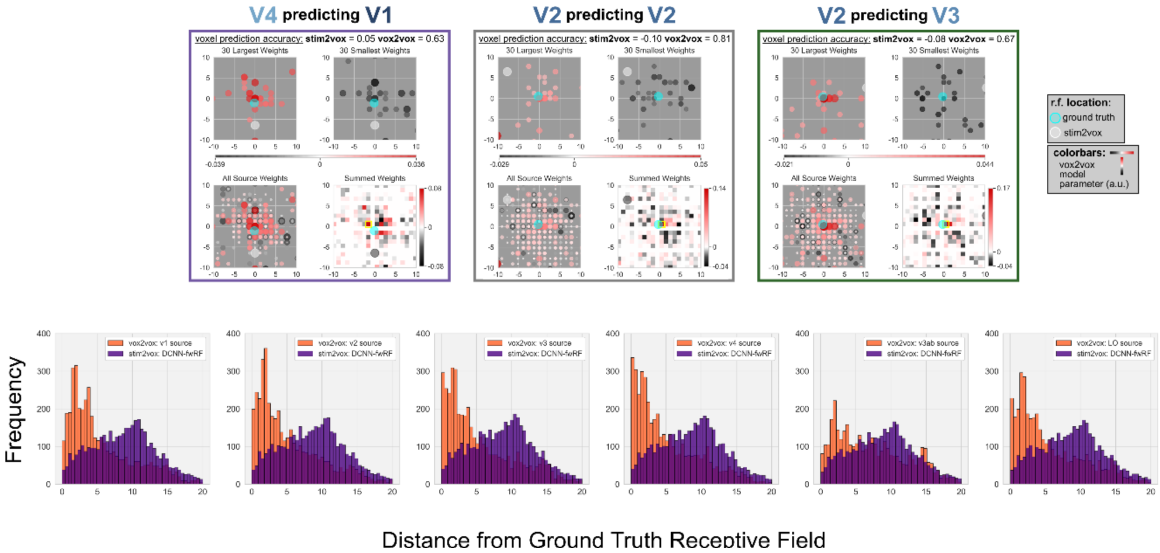

Figure 4.2 *Estimating receptive field location from voxel-to-voxel model parameters.* **Top:** The source voxel weights of voxel-to-voxel models for three target voxels. For each example voxel the prediction accuracy of the stimulus-to-voxel model fell below significance threshold; the prediction accuracy of the voxel-to-voxel model was above threshold. When plotted in visual space (gray panels) according to the receptive field locations of the source voxels, the largest positive voxel-to-voxel model weights (circles in top left of each box; circle radius and intensity scale with magnitude of weight) cluster near the "ground truth" receptive field location of the target voxel (aqua circle; estimated from an independent retinotopic mapping experiment). The largest negative weights (top right of each box) tend to cluster in the near periphery of the ground-truth receptive field location. For these voxels, the receptive field location estimated from the stimulus-to-voxel model (white circle) is misplaced relative to the location of the ground-truth receptive field. Visualizations of all weights (bottom left of each box) and sums of weights for each receptive field location (bottom right) also reveal distinct peaks of positive weight values near the ground-truth receptive field location. The receptive field location estimated from the voxel-to-voxel model weights is the location with the maximum sum of source voxel weights. **Bottom:** The distance (in degrees of visual angle) between the "ground truth" receptive field location and the locations estimated from the stimulus-to-voxel (purple bars) and voxel-to-voxel (coral bars) models is calculated for target voxels that have a stimulus-to-voxel prediction accuracy below the significance threshold (i.e., voxels in the "ball of nothingness"). Histograms of these distances for source area V1 (leftmost panel) through source area LO (rightmost) show that receptive field locations estimated from voxel-to-voxel models are generally closer to the ground truth receptive field locations than those estimated from stimulus-to-voxel models.

To quantitatively assess this clustering we estimated a ‘voxel-to-voxel receptive field’ location. This location was calculated using a weighted 2-dimensional histogram in which each X,Y location in the stimulus-to-voxel fwRF candidate grid is one ‘bin’ in the 2d histogram. We extracted the weight parameters for source voxels above a 0.2 stimulus-to-voxel fwRF prediction accuracy threshold in a target voxel's model and binned them according to their corresponding X,Y coordinate estimated by the stimulus-to-voxel fwRF. The voxel-to-voxel receptive field is the X,Y location corresponding to the bin with the maximum sum of weights (**Figure 4.2** Top, bottom right visual field plots).

We then calculated the Euclidean distance between the 'ground truth' receptive field (as estimated via pRF analysis of a separate retinotopic mapping experiment) and both the stimulus-to-voxel fwRF and voxel-to-voxel receptive fields. Although estimates of receptive field location derived from voxel-to-voxel models were most accurate(closest to ground truth) when the source and target voxels belonged to the same visual area, estimates were more accurate than receptive field locations derived from the stimulus-to-voxel fwRF model even for hierarchically distant source-target pairings (**Figure 4.2** Bottom). Thus, for a given target voxel the source of variance unexplained by the stimulus-to-voxel models during natural scene stimulation is not shared by all voxels in the same brain, but is shared with (and only with) voxels that have overlapping receptive field locations (i.e., voxels that co-activate during retinotopic mapping stimulation).

4.3.3 Prediction accuracy of voxel-to-voxel models depends on signed, hierarchical distance between source and target

The relationships between patterns of activity (and the representations those patterns encode) in distinct visual areas in the brain are undoubtedly nonlinear. Intuitively, the relationships between source and target voxels in different brain areas should therefore show some resistance to linear voxel-to-voxel modeling. We might expect this resistance to be especially strong for hierarchically distant brain areas that are known to encode stimuli into very different visual features. Thus, we examined median prediction accuracy of the voxel-to-voxel models for each pairing of source and target visual area as a function of hierarchical distance and sign.

Consistent with our expectations, we found that median prediction accuracy for any target area was highest for lateral models (i.e., source voxels in same area as area of target voxel) but then declined monotonically as hierarchical distance between a source and target area increased in the feed-forward direction.

Yet several aspects of the relationship between source and target areas were somewhat unexpected. The prediction accuracy of feedback models did *not* decline with hierarchical distance (**Figure 4.3** A & B) between source and target area, and was higher than the prediction accuracy of the feed-forward model (**Figure 4.3** D) for most source/target pairs. Finally, while the lateral model was most accurate for each target area, median prediction accuracy for lateral models declined with ascension of the visual hierarchy (**Figure 4.3** C).

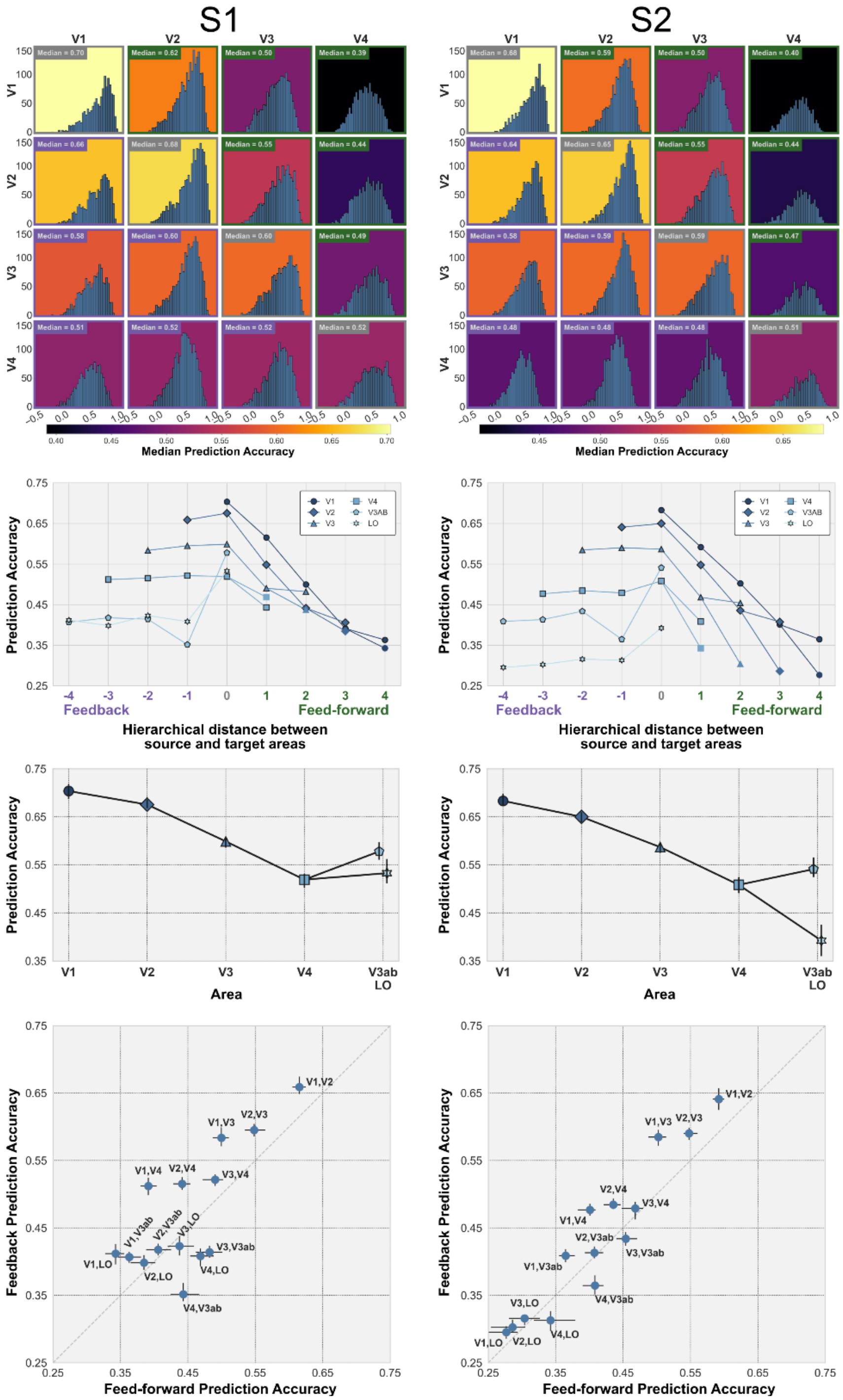

Figure 4.3 *Patterns of prediction accuracy*. **A:** Distribution (voxel count on y-axis) of prediction accuracy (x-axis; background color indicates median of distribution) for voxel-to-voxel model source (row) target (column) pairings. **B:** Median prediction accuracy (y-axis) of feed-forward models declines with hierarchical distance (x-axis; 0 = lateral model) between source (indicated by color of each curve) and target (indicated by distance to source). Median prediction accuracy of feedback models not dependent on hierarchical distance. Areas V3ab and LO same level in the hierarchy, LO targets have grey border. **C:** Median prediction accuracy of lateral models decreases with hierarchical position. **D:** Median prediction accuracy of feedback models (y-axis) is larger than median prediction accuracy of feed-forward models (x-axis) for most pairs of visual brain areas (blue dots).

4.3.4 Single Trial Analyses

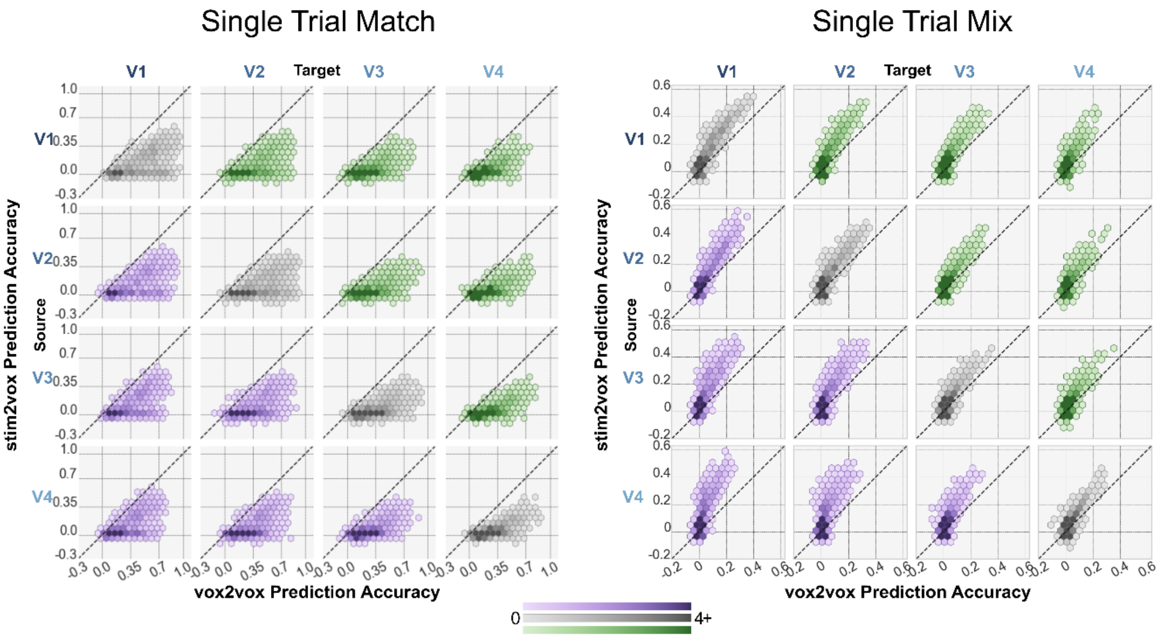

Figure 4.4 *Voxel-to-voxel model prediction accuracy for single trial analyses*. Format as in Figure 4.1. **Left:** Single-trial analyses where source and target activity are from the same trial out-perform stimulus-to-voxel fwRF single-trial models, similar to repeated trial analyses. **Right:** Single-trial analyses where source and target activity are from different trials perform about as well as single-trial stimulus-to-voxel fwRF, similar to cross-subject analyses. All data from S1.

Repeated-trial analyses are generally performed in order to average out ‘noise’ associated with trial-to-trial variability. However, this ‘noise’ may in fact be unexplained signal variance. Therefore, we applied voxel-to-voxel models to single trial data in two ways. First, in the same manner as the repeated-trial analysis, source and target activity were matched trial to trial. Second, we ‘mixed’ the trials such that source activity from trial 1 was trained to predict source activity to trial 2 and vice versa. This ensured the same number of training trials in both models, however the mix model allows us to examine if the variance explained by voxel-to-voxel models is stable across trials.

Like repeated-trial models, matched single-trial data (**Figure 4.4** left) show predictive gains in voxel-to-voxel models. These gains appear to be bigger than in repeated-trial measures, perhaps owing to the voxel-to-voxel models ability to utilize signal previously considered to be noise in stimulus-to-voxel models. Conversely, mixed single-trial models (**Figure 4.4**, right) are more similar to cross-subject models in that they can only predict about as well as the stimulus-to-voxel model. This result may indicate trial-to-trial variability stems from transient rather than stable, on-going activity. However, this would need more investigation with an experimental design more suited to investigating trial-to-trial variability over time.

4.3.5 Control Analyses

To ensure voxel-to-voxel models are exploiting meaningful signal we performed several control analyses. First, a bootstrap resampling of stimuli (**Figure 4.5**), shows our results are robust with little variance across bootstraps. Within-subject voxel-to-voxel models consistently out-perform stimulus-to-voxel models across bootstraps, while cross-subject models consistently hover around the line of parity.

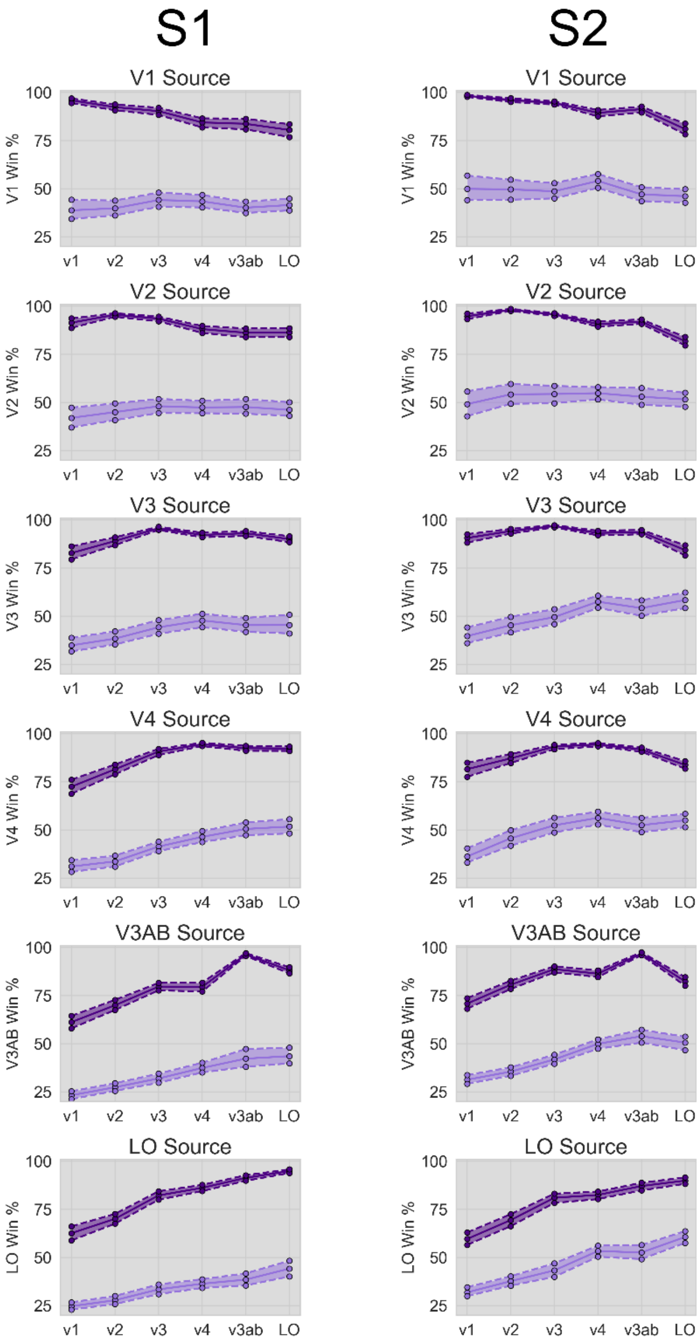

Figure 4.5 *Prediction accuracy of voxel-to-voxel models relative to stimulus-to-voxel models.* The win percentage (y-axis) is the percentage of voxels for which the voxel-to-voxel model has a higher prediction accuracy than the stimulus-to-voxel model. A win percentage at 50% indicates that voxel-to-voxel and stimulus-to-voxel models have roughly equal prediction accuracy across a population of voxels. When voxel-to-voxel models are estimated for source and target voxels in the same brain (dark purple curves; dashed line indicates median; shading indicates 5th/95th percentile over 1,000 bootstrapped samples of trials), win percentages exceed 50% for all source (sub-panel titles) and target (x-axis) pairs. When voxel-to-voxel models are estimated for source and target voxels in the different brains (light purple Curves), win percentages are at or below 50% for all source and target pairs.}

Next, we assessed how different numbers of source voxels would affect the hierarchical prediction accuracy patterns in voxel-to-voxel models. We applied voxel-to-voxel models with a fixed number of voxels (N=100) per source region to determine if the asymmetry in prediction accuracy was due to smaller ROIs higher in the hierarchy. However, the predictive advantage and asymmetry was still present with equal numbers of source voxels per area (**Figure 4.6**, left). Additionally, we applied voxel-to-voxel models with randomly selecting 20% of each ROI's voxels as the source input. This again did not change our results. Both analyses are particularly relevant to lateral models. Lateral models are most susceptible to spatially correlated noise, with the possibility that voxels directly adjacent to the target voxel are solely responsible for the increase in prediction accuracy. However, randomly selecting voxels mitigates the effects of spatially autocorrelated noise to an extent.

Due to spatially autcorrelated noise factors in fMRI data, we expect there to be some increase in prediction accuracy for voxel-to-voxel over stimulus-to-voxel models. Yet the structure revealed via voxel-to-voxel receptive field analysis and the increase in prediction accuracy for hierarchically distant source-target pairings imply this advantage is not entirely due to noise. Disentangling the specific source(s) of structured, unexplained variance is not possible in the context of this dataset. However, to determine the effects of removing noise and stimulus related signal on voxel-to-voxel models, we performed four additional 'control' analyses.

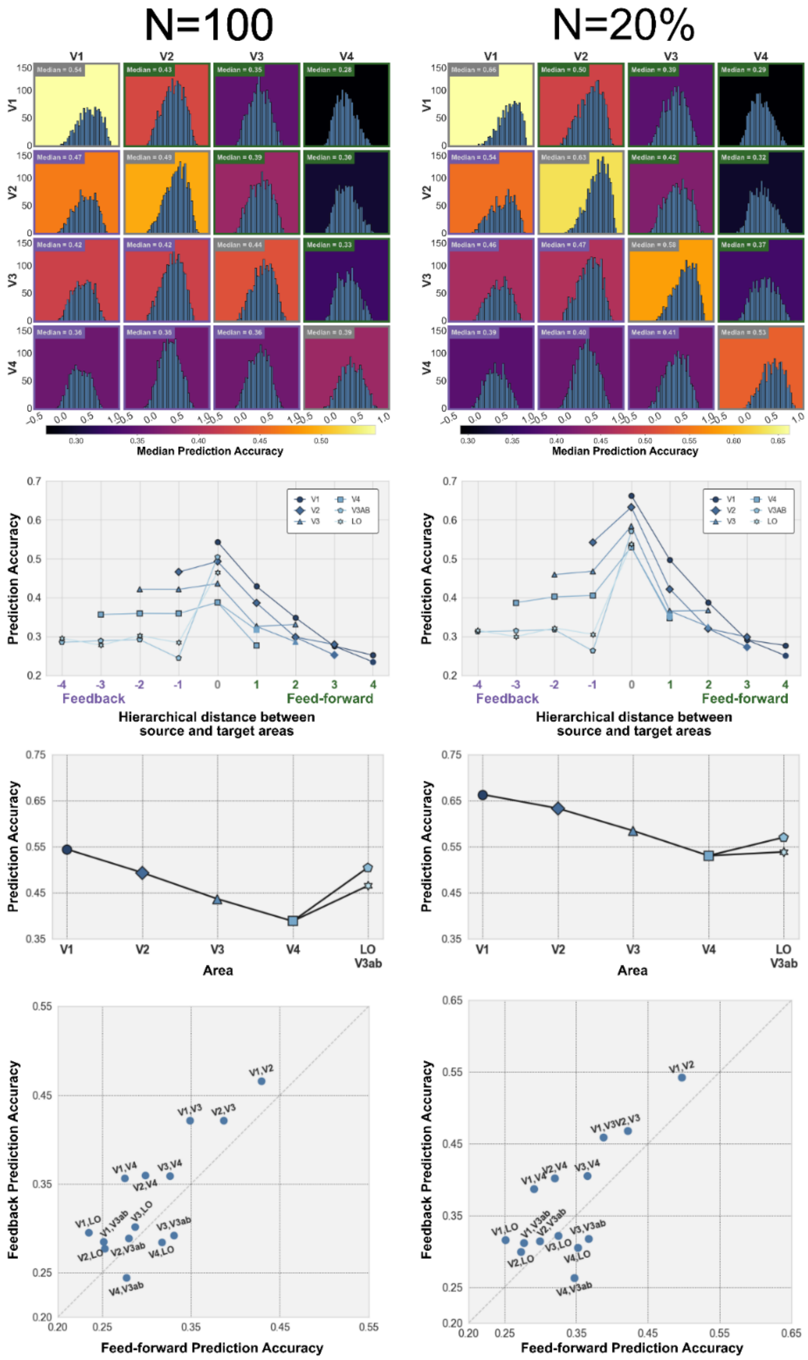

Figure 4.6 *Voxel-to-voxel model prediction accuracy for fixed numbers and percentages of source voxels*. Format as in Figure 5.3 **Left:** To control for variation in the number of voxels across source areas we randomly sampled 100 voxels from each source area then re-estimated voxel-to-voxel models. We report the average of the median prediction accuracy across 10 random samples for each source-target pairing. **Right:** Here we randomly sample 20% of voxels in each source area. All data from S1

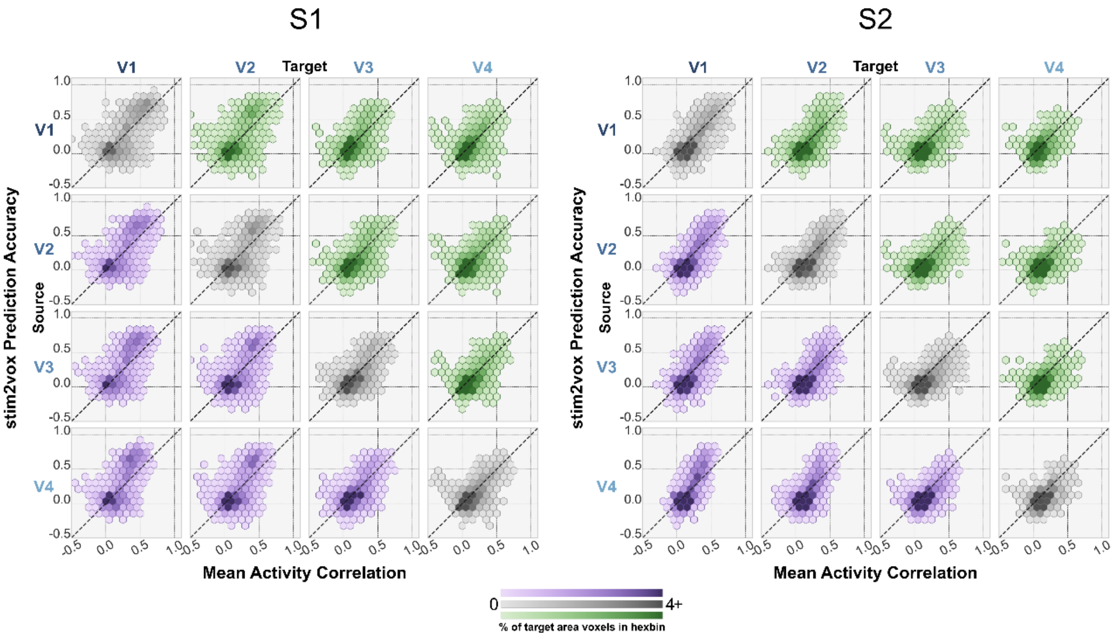

Figure 4.7 *Averaging out spatial structure eliminates predictive advantage*. Each target voxel's activity was correlated with the mean activity for every source area. Simply correlating with mean activity does not provide advantage of the prediction accuracy of the stimulus-to-voxel fwRF model.

First, as a baseline measure, we correlated each target voxel's activity with the mean activity of each source area. If correlated noise was the sole source of unexplained variance, averaging out spatial structure would still yield correlations higher than the stimulus-to-voxel fwRF prediction accuracies. However mean activity correlations for source areas did not outperform predictions from the stimulus-to-voxel fwRF model (**Figure 4.7**).

Next, we created two additional source areas. The first consisted of voxels selected as far from brain voxels as possible, we refer to this as an 'Air' source area. Voxels from white matter were selected as the second source area. We then subtracted the predictions made by each new source, separately, from the original beta activation values. If scanner related noise is the main component of unexplained variance, subtracting predictions from the Air source area should negate the advantage from our original models. Similarly, white matter would be subject to the same physiological-based noise as brain voxels, therefore if physiological noise is the main source of unexplained variance, removing white matter predictions would negatively impact all voxel-to-voxel models. We applied voxel-to-voxel modelling to the new adjusted values and found neither had an appreciable effect on voxel-to-voxel prediction accuracy (**Figure 4.8**).

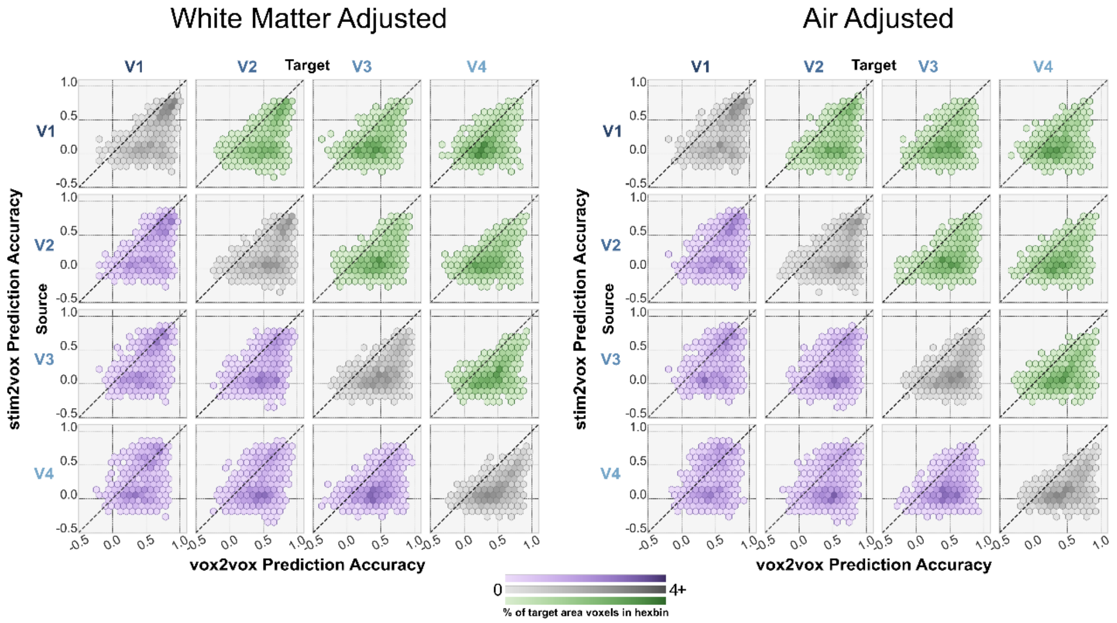

Figure 4.8 *Removing activity from noise sources does not degrade voxel-to-voxel predictive advantage*. Two additional source areas were created, one consisting of white matter voxels and one from 'air' voxels far from the brain. Voxel-to-voxel predictions made from those source areas were then subtracted from the original beta activation values. Voxel-to-voxel models for the original source-target pairing were then applied to the new adjusted values. White-Matter adjusted voxel-to-voxel models on left, air-adjusted models on right. Neither 'noise' source contributes significantly to the voxel-to-voxel model predictive advantage.

Finally, we applied voxel-to-voxel models to the residuals of the stimulus-to-voxel fwRF model predictions. We again calculated voxel-to-voxel receptive field locations and distance from ground truth receptive field. The voxel-to-voxel receptive field locations remained closer to ground truth than the stimulus-to-voxel fwRF, even for voxels well predicted by the stimulus-to-voxel fwRF (**Figure 4.9**). This is further evidence that unexplained variance is structured and respects retinotopic principles without relying on stimulus-based signal.

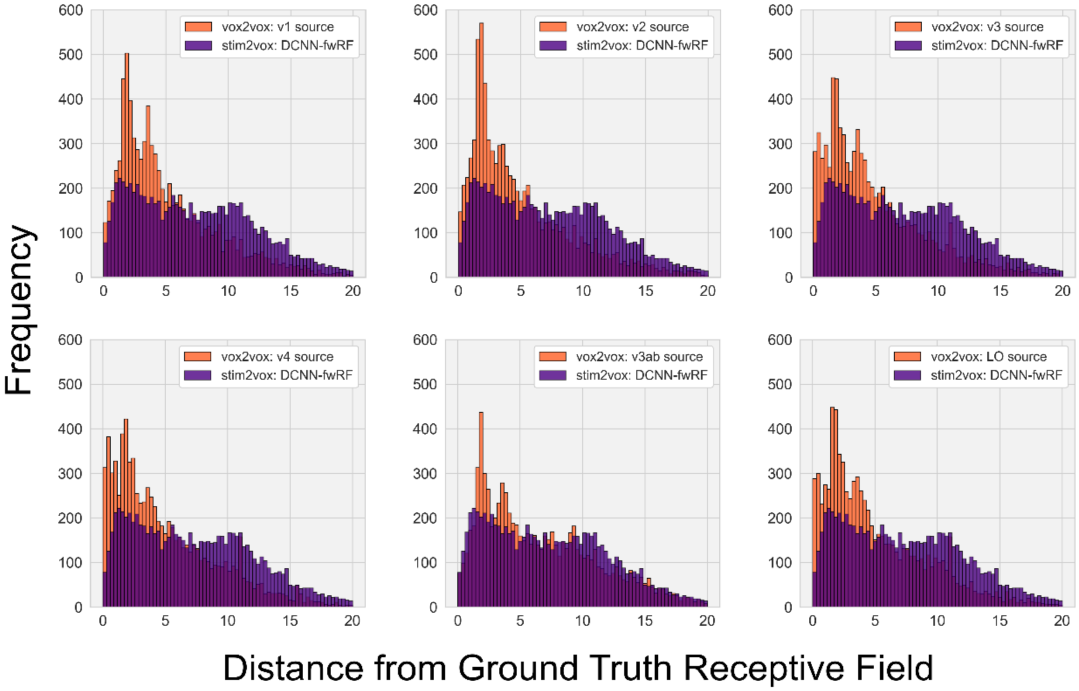

Figure 4.9 *Voxel-to-voxel receptive field locations are closer to ground truth, even after removing stimulus-based activity*. Voxel-to-voxel models were applied to the residuals of the stimulus-to-voxel fwRF model. Voxel-to-voxel receptive field locations and their Euclidean distance from ground truth receptive fields, as estimated by a separate retinotopic mapping experiment, were recorded. Each histogram plots all target voxels' distance values as predicted by the indicated source area, as compared to the values predicted by the stimulus-to-voxel fwRF. Although the stimulus-based activity, as predicted by the stimulus-to-voxel fwRF, has been removed, voxel-to-voxel models still estimate receptive field locations closer to ground truth across all target voxels and eccentricities.

4.4 Conclusion

A linear transformation of activity in source voxels predicted activity in nearly every target voxel more accurately than an optimized, nonlinear transformation of the stimulus. This finding clearly demonstrates that the stimulus-to-voxel model is blind to one or more ``hidden" sources of variance that induce strong correlations between the activities of voxels across the visual hierarchy. These hidden sources of variance must be endogenous (i.e., not entirely stimulus-dependent) because the voxel-to-voxel model did not predict more accurately than the stimulus-to-voxel model when source and target voxels were located in different brains (**Figure 4.1** Right). Importantly, we have shown that the correlations induced by these hidden, endogenous sources of variance are highly structured and appear to be dependent upon representations encoded in the brain activity. Induced correlations are strongest between voxels with adjacent receptive fields, even when source and target voxels are hierarchically distant (**Figure 4.2**) and when stimulus-related signal is removed (**Figure 4.9**). The extent to which linear voxel-to-voxel models can exploit induced correlations to achieve accurate predictions depends upon the hierarchical locations of the source and target voxels (**Figure 4.3** A & B). The prediction accuracy of lateral voxel-to-voxel models degrades with ascent of the visual hierarchy (**Figure 4.3** C), and the prediction accuracy of feedback models is larger than the corresponding feed-forward models for most source/target pairs (**Figure 4.3** D). Mean activity correlations and predictions from Air or White Matter voxels cannot account for the increase in prediction accuracy of voxel-to-voxel models (**Figure 4.8**). Finally, stimulus-based signal cannot account for the decrease in distance to ground truth receptive field locations in the voxel-to-voxel receptive field model estimates.

Chapter 5: Pixel-to-Pixel Models

Specific Aim 3: *Build pixel-to-pixel encoding models of unit activations from deep neural networks trained for object recognition and compare results to Aims 1 2.* Hypothesis: The pattern of predictive connectivity will be different from that seen in the hierarchy of brain areas.

5.1 Overview and Rationale

The differences in patterns between artificial and real neural networks may reveal important insights into how to better structure future artificial neural nets to more closely resemble computations made by the brain. To provide recommendations for future stimulus-to-voxel models, we need to be able to examine the underlying architecture of these models in a similar way to how we examine the brain. Therefore, using the same principles of voxel-to-voxel models, we developed 'pixel-to-pixel' models utilizing unit activations in different layers of two DCNN's trained for object recognition

The first DCNN is one built and trained in-house for object recognition on the CIFAR-10 Dataset, this DCNN was chosen because the structure approximates the AlexNet DCNN architecture, without as many parameters. This allows for complete pixel-to-pixel models. Additionally, we applied pixel-to-pixel models to AlexNet, the DCNN underlying the stimulus-to-voxel fwRf model, but due to computational constraints we could only model a randomly selected subset of pixels in each layer.

5.2 Methods

As with voxel-to-voxel models we fit a pixel-to-pixel model for every possible pair of source layer and target layer, and refer to pixel-to-pixel models as feedforward, feedback and lateral depending upon the relative positions of the source and target layers in the network hierarchy. Note, due to computational constraints, lateral models were calculated slightly differently for pixel-to-pixel models. Specifically, 10% of pixels in a layer were randomly selected as target pixels, with the remaining 90% of pixels as the source pixels. This procedure was repeated ten times such that a lateral model was computed for every pixel in a layer. Due to the high redundancy of feature information in DCNN layers, we do not expect this procedure to affect the lateral model prediction accuracy. With the AlexNet DCNN we randomly subsampled 10% of each layers pixels and used those subsamples in each source-target pairing.

5.3 Results

A very different relationship between prediction accuracy and hierarchical location was observed when we estimated linear approximations to the connections between layers in the DCNNs. As in the brain, in the CIFAR10 DCNN median prediction accuracy for any target node was highest for lateral models (i.e., source nodes in same layer as the layer of the target node), and median prediction accuracy declined monotonically as hierarchical distance between a source and target layer increased. In contrast to the brain, the prediction accuracy of *feedback* pixel-to-pixel models declined *more* rapidly with hierarchical distance between layers in the

CIFAR10 DCNN than the feed-forward pixel-to-pixel models, and was *lower* than the prediction accuracy of the feed-forward model for all source/target pairs. Finally, median prediction accuracy for lateral models *increased* with ascension of the network hierarchy (**Figure 5.1** left).

Pixel-to-pixel models built on AlexNet DCNN layers provided a different pattern from both the brain and the CIFAR-10 DCNN (**Figure 5.1** right). For instance, Layer 2 target pixels are poorly predicted, regardless of source layer. Even in lateral models, Layer 2 has the lowest prediction accuracy. Unlike the CIFAR-10, but similar to the brain, there is a sharp decline in feed-forward models, perhaps owing to AlexNet's architecture resembling the bottom up processing in the visual system. There is some asymmetry in feedback versus feedforward models. Excluding models where L2 is a target, all other models prefer the feedback direction, a stark difference from the CIFAR-10 network and again similar to the brain. In general, pixel-to-pixel models applied to AlexNet layers perform worse than pixel-to-pixel models in the CIFAR-10 network, however it is unclear if this is due to the subsampling procedure necessary for AlexNet models.

5.4 Conclusion

Both AlexNet and the CIFAR-10 DCNNs lack the lateral and feedback connections present in the brain. These connections no doubt contribute to the brain's ability to share information across wide regions of cortex. Therefore, it is unsurprising that the patterns of prediction accuracy seen between brain ROIs is not replicated between the layers of the DCNNs. The slight similarities between AlexNet and the brain suggest that we have successfully modelled one aspect of visual computation, the feedforward information flow. However, this may imply that we have reached the limit at which we can explain brain data with solely stimulus-based information.

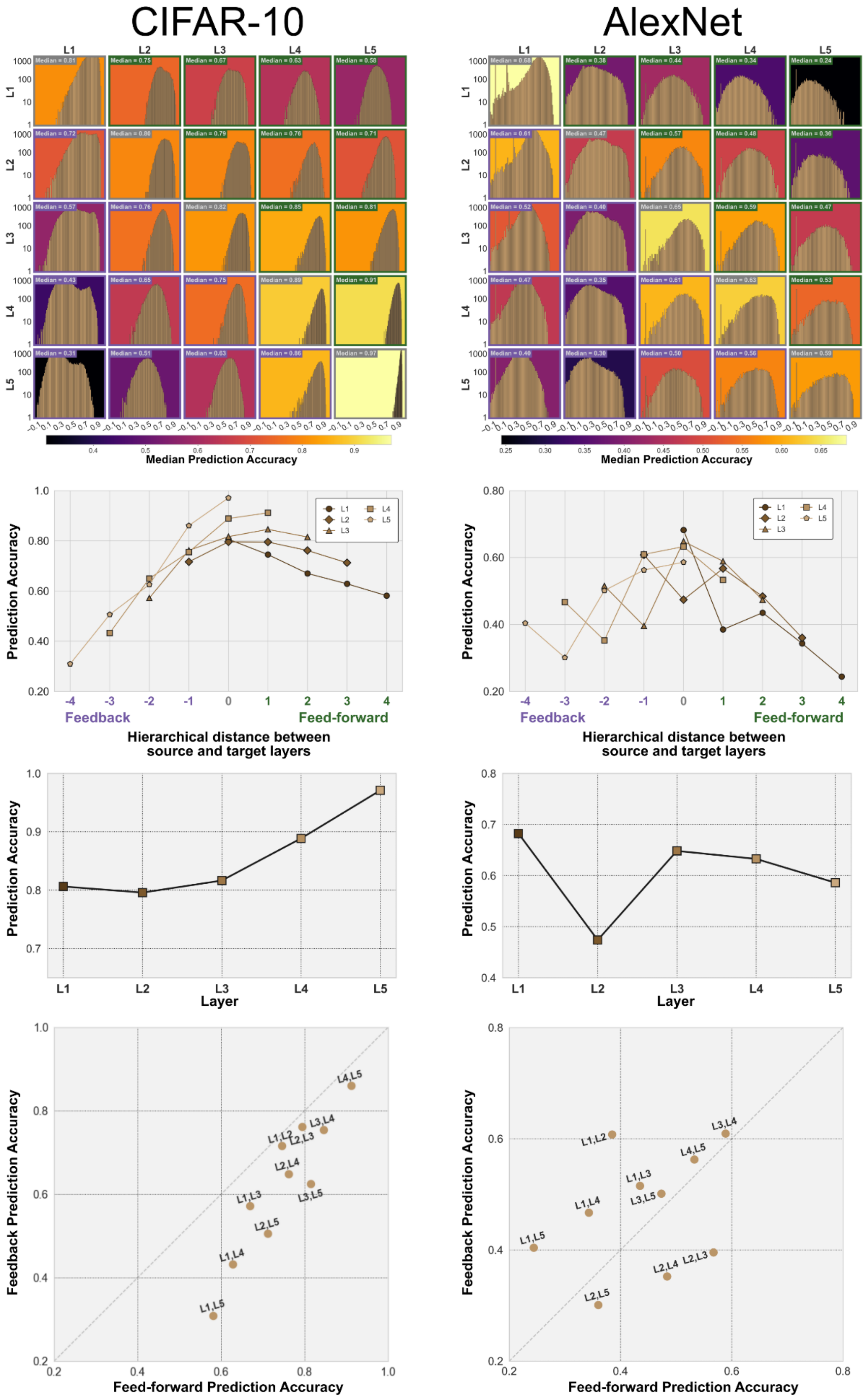
CIFAR-10
AlexNet
L1
L2
L3
L4
L5
Median = 0.81
Median = 0.75
Median = 0.67
Median = 0.63
Median = 0.58
Median = 0.72
Median = 0.80
Median = 0.79
Median = 0.76
Median = 0.71
Median = 0.57
Median = 0.76
Median = 0.82
Median = 0.85
Median = 0.81
Median = 0.43
Median = 0.65
Median = 0.75
Median = 0.89
Median = 0.91
Median = 0.31
Median = 0.51
Median = 0.63
Median = 0.86
Median = 0.97
Median = 0.68
Median = 0.38
Median = 0.44
Median = 0.34
Median = 0.24
Median = 0.61
Median = 0.47
Median = 0.57
Median = 0.48
Median = 0.36
Median = 0.52
Median = 0.40
Median = 0.65
Median = 0.59
Median = 0.47
Median = 0.47
Median = 0.35
Median = 0.61
Median = 0.63
Median = 0.53
Median = 0.40
Median = 0.30
Median = 0.50
Median = 0.56
Median = 0.59
Median Prediction Accuracy
Prediction Accuracy
Feedback
Feed-forward
Hierarchical distance between source and target layers
Layer
Feedback Prediction Accuracy
Feed-forward Prediction Accuracy
L4,L5
L1,L2
L3,L4
L2,L3
L2,L4
L1,L3
L3,L5
L2,L5
L1,L4
L1,L5

Figure 5.1: *Patterns of prediction accuracy across layers of deep neural networks*. A: Sub-panels show the distribution (pixel count on y-axis) of prediction accuracy (x-axis; background color indicates median of distribution) for pixel-to-pixel models of CIFAR-10 (left panels) and AlexNet (right panel). Sources(rows) and targets(columns) are layers numbered from L1 (closest to input) to L5(farthest from input). B. In the CIFAR-10 network median prediction accuracy of feed-forward pixel-to-pixel models declines slowly with hierarchical distance; median predication accuracy of feedback models declines more rapidly. However in AlexNet, the decline in feed-foward models is not as steep. Interestingly, regardless of source layer, prediction accuracy values for Layer 2 are low. C: Median prediction accuracy of lateral pixel-to-pixel models in CIFAR=10 increases with hierarchical position of source and target layer, in AlexNet there is a shallow decline, save for the sharp decrease in Layer 2. D: In CIFAR-10 median prediction accuracy is higher for feed-forward versus feedback models for each pair of network layers (brown dots) The reverse is true in AlexNet, with the exception of pairings where L2 is the target.

Chapter 6: Neuron-to-Neuron Models

6.1 Overview and Rationale

In the previous chapters we have provided evidence that unexplained variance in natural scene fMRI data is a structured, functionally relevant signal. However, the signal from an fMRI voxel is an indirect vasculature-based measure that potentially reflects the summation of activity over thousands of neurons. Therefore, it is unclear if the ability of voxel-to-voxel models to capitalize on correlated activity in the brain is due the spatial scale and method of measurement or reflective of intrinsic neural activity the neuron level. In this chapter we apply our methodology to a publicly available 2-Photon dataset. Activity from approximately 10,000 V1 neurons was recorded while mice passively viewed natural scene images. Similar to voxel-to-voxel lateral models, our neuron-to-neuron model will predict activity for a target neuron based on activity of all other recorded neurons.

6.2 Methods

Neural activity in V1 was recorded in mice bred to express GCaMP6s. The mouse we chose for our analysis also expressed tdTomato, allowing for identification of excitatory vs inhibitory neurons. Full details regarding the experimental procedures can be found in (Stringer et al., 2019b).

We applied the two versions of feature-weighted Receptive Field (fwRF) model developed by our lab to estimate receptive field locations for each neuron.

One model was based on Gabor-wavelets and one based on feature maps extracted from the DCNN AlexNet, both used the same candidate receptive field grid. In the original experiment the visual stimulus subtended 270 x 68 degrees of visual angle. To reduce computation time we restricted the candidate grid to cover from – 104 degrees to 14 degrees in the horizontal plane of the visual field, the entire vertical span was included. We chose these values based on the receptive field locations estimated in the original paper. Candidate centers were linearly spaced in both the X (N=56) and Y (N=35) directions. 4 log-spaced size parameters between 3 and 12 degrees were considered, for a total of 7840 candidate models.

For the gabor-fwRF, we generated 56 Gabor wavelets at 7 linearly spaced spatial frequencies between .01 and .13 cycles per degree. Each frequency sampled 8 evenly spaced orientations between 0 and π. For the DCNN fwRF we followed the same procedure detailed in Chapter 2 for voxel-based models. Briefly, we fed all natural scene stimuli from the 2-Phton experiment through AlexNet and extracted the resulting feature maps.

6.3 Results

6.3.1 Validation of stimulus-to-neuron fwRF encoding models on 2-photon data

Prior to applying neuron-to-neuron models, we must first validate the stimulus-to-neuron fwRF encoding models on the 2-photon data. **Figure 6.1** (left) compares the distributions of prediction accuracy for both types of fwRF encoding models (Gabor-based and DCNN-based). Both models successfully predict receptive field location for many neurons, with generally higher prediction accuracies than those obtained with fMRI data. Similar to fMRI, the DNN-based model has a predictive advantage over the Gabor-based.

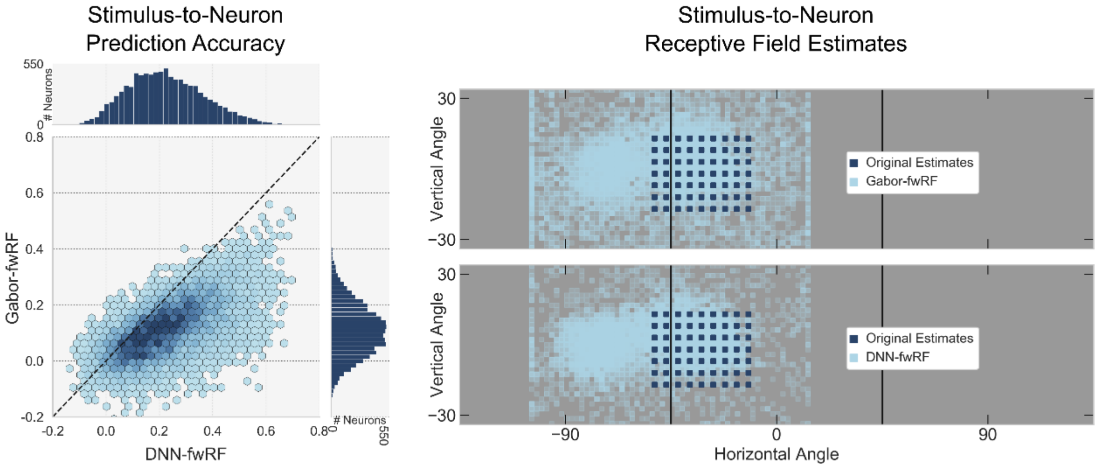

Figure 6.1 *Stimulus based feature weight receptive field model can accurately predict receptive field locations for neurons in mouse V1.* **Left:** Both Gabor-fwRF (prediction accuracy y-axis) and DCNN-fwRF (prediction accuracy on x-axis) predict 2-photon imaging data with high accuracy for many neurons. DCNN-fwRF shows a slight advantage over Gabor. **Right:** Both models (Gabor on top, DCNN on bottom) predict receptive field locations overlapping with the original paper estimate. However due to the computational flexibility of the fwRF we were able to extend our grid much further and reveal many neurons prefer receptive field location further left in the visual field.

Visual space plots show predicted receptive field location for both models as compared to the locations predicted by the original data paper (**Figure 6.1**, right). The authors used a coarse-to-fine approach which narrowed down the candidate receptive fields for all neurons to the 9x7 grid indicated in dark blue. Both fwRF models predict receptive fields within a similar area of visual space, however many estimates extend further left into the visual field than predicted by the original model. The original researchers limited their grid mainly due to computational restraints, but it appears the fwRF model can estimate receptive fields more efficiently over a larger grid. The combination of high prediction accuracy values and significant overlap with the original estimates of receptive field location indicates the fwRF is an accurate and valid method for estimating receptive field locations using 2-Photon data. For the rest of the analyses we will use the AlexNet-based stimulus-to-neuron fwRF as our reference model.

6.3.2 Comparison of neuron-to-neuron encoding models with stimulus-to-neuron encoding model

We next applied the same approach used in voxel-to-voxel and pixel-to-pixel models to the mouse data. Figure 7.2 shows the results of this modeling procedure for all neurons in one mouse with neuron-to-neuron model prediction accuracy on the x-axis of each plot and DCNN-fwRF prediction accuracy on the y-axes. On the left all neurons are plotted and the pattern seen is identical to the patterns seen in the lateral models of voxel-to-voxel models. The middle and right plots separate excitatory from inhibitory cells. The pattern in excitatory cells remains similar with

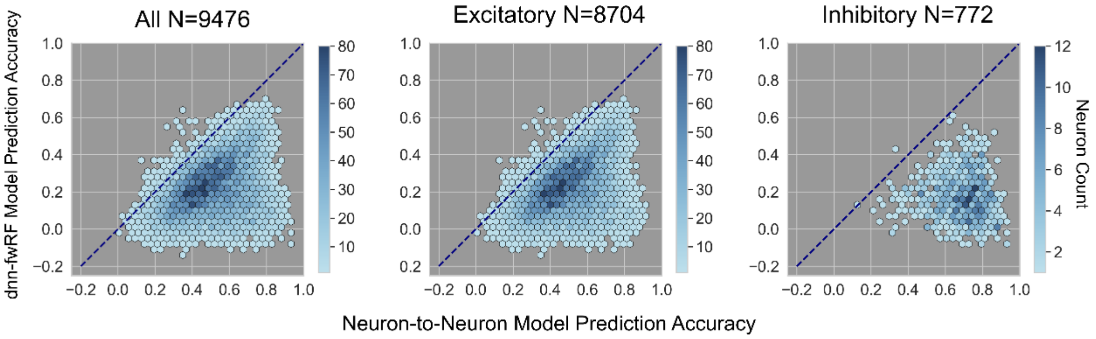

Figure 6.2 *Neuron-to-neuron models out predict DCNN-fwRF.* **Left:** Distribution of prediction accuracy for all neurons for DCNN-fwRF (y-axis) and neuron-to-neuron (x-axis). Similar pattern to voxel-to-voxel lateral model emerges. **Middle:** Same as left but restricted to neurons labeled as excitatory. **Right:** Same as other but restricted to neurons labeled as Inhibitory. Intriguingly these neurons appear to enjoy the largest benefit from neuron-to-neuron modeling.

the majority of neurons showing a slight predictive advantage in the neuron-to-neuron model. The inhibitory cells also show an advantage, however it appears to be more pronounced than that of many excitatory cells.

We also applied the DNN-fwRF and neuron-to-neuron models to single trial mouse data. Similar to voxel-to-voxel analysis, neuron-to-neuron models were

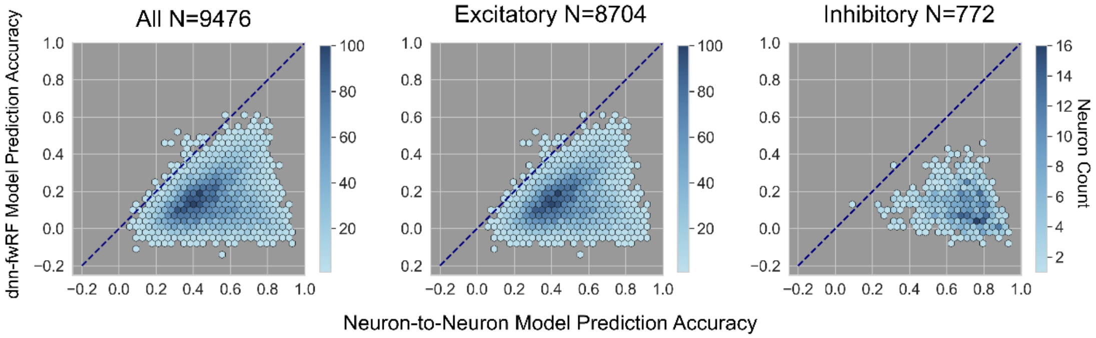

Figure 6.3 *Neuron-to-neuron models out predict DCNN-fwRF in single trial data.* **Left:** Distribution of prediction accuracy for all neurons for DCNN-fwRF (y-axis) and neuron-to-neuron (x-axis). Similar pattern to repeated trial data. **Middle:** Same as left but restricted to neurons labeled as excitatory. **Right:** Same as other but restricted to neurons labeled as Inhibitory. Again these neurons appear to enjoy the largest benefit from neuron-to-neuron modeling.

applied in two ways. In the match conditions source and target neurons were drawn from the same trials, in the mix condition source neurons were trained to predict target neurons on a different trial of the same stimulus. Once again we see the same pattern as voxel-to-voxel models. The single-trial match models show a predictive advantage over the dnn-fwRF for all cell types, just as in the repeated trial data. The single-trial mix models are predominantly at par with the dnn-fwRF model, however there does appear to be some neurons, particularly those identified as inhibitory cells, that do still see an increase in prediction accuracy in the neuron-to-neuron models.

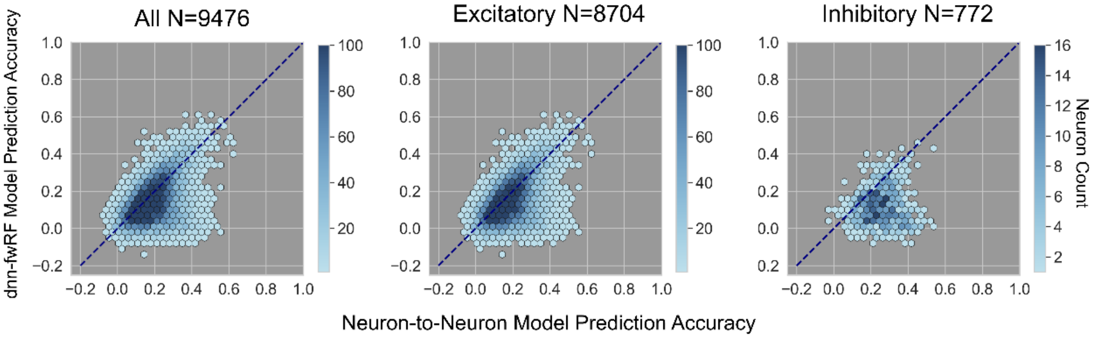

Figure 6.4 *Neuron-to-neuron models at par with DCNN-fwRF in single trial mix analysis.* **Left:** Distribution of prediction accuracy for all neurons for DCNN-fwRF (y-axis) and neuron-to-neuron (x-axis). Similar pattern to voxel-to-voxel single trial mix models emerges. The majority of neurons are at par with stimulus based predictions, with a small group of neurons seeing improvement. **Middle:** Same as left but restricted to neurons labeled as excitatory. **Right:** Same as other but restricted to neurons labeled as Inhibitory. Intriguingly these neurons appear to those that benefit from neuron-to-neuron modeling in mixed trials.

6.4 Preliminary Results & Conclusion

Replicating results across species and imaging modalities is an important and exciting step in determining the sources and role of intrinsic neural activity. Using the neuron-to-neuron method in 2-photon data opens a wide range of experimental paradigms. One interesting addition to our results comes from a spontaneous

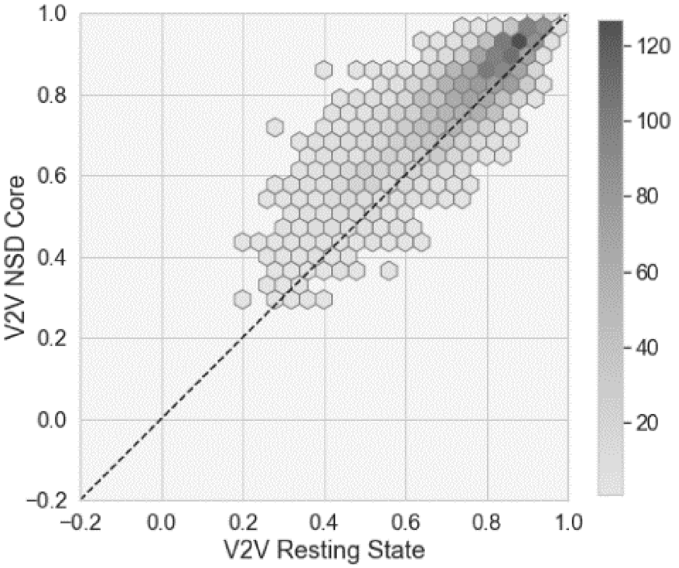

Figure 6.5 *Voxel-to-voxel models trained on natural scenes successfully predict resting state data.* Lateral V1 voxel-to-voxel model built on natural scenes data can predict resting state data (prediction accuracy x-axis) with the same accuracy as natural scene images. (prediction accuracy y-axis) for almost all voxels in human V1.

activity experiment performed with mice in the original dataset. Spontaneous activity is naturally intrinsic and potentially comparable to resting state activity recorded in human fMRI experiments. While the vim-1 dataset does not include resting state, a new unreleased natural scenes dataset recently acquired by our group does. We present preliminary results relating spontaneous activity prediction patterns in mice and humans. We built voxel-to-voxel models in the same exact method as previously described with the new dataset. We then used the resulting models to predict resting state data acquired from the same subject. **Figure 6.3** compares prediction accuracy for natural scene images vs resting state time series in human V1. For most voxels, the voxel-to-voxel model trained on natural scene data can predict resting state time series just as well as held out natural scene activity. This indicates the intrinsic activity during task paradigms may be related to spontaneous activity captured during resting state scans.

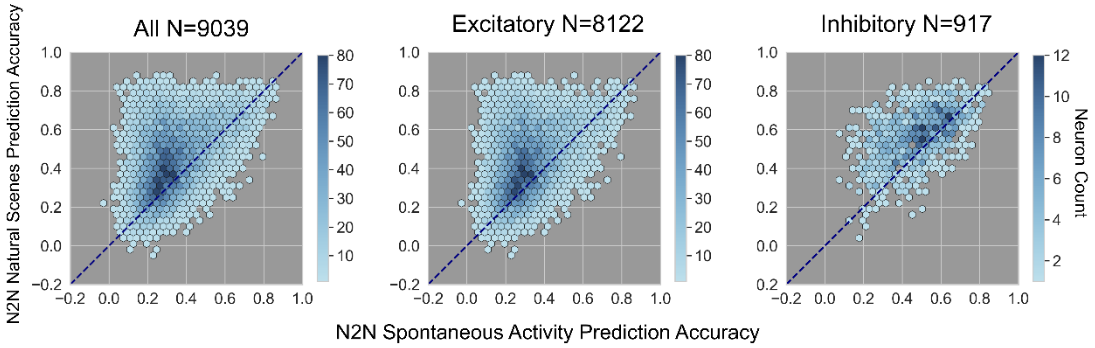

Figure 6.6 *Neuron-to-neuron models trained on natural scenes successfully predict spontaneous activity data.* **Left:** Neuron-to-neuron models built on natural scene images (prediction accuracy y-axis) predict spontaneous activity with similar accuracy in almost all neurons. **Middle, Right:** Same as left but split into excitatory and inhibitory neurons. It appears the group of neurons that are better predicted with natural scenes data are mostly excitatory.

We then performed this same analysis in the 2-photon mouse data. We built neuron-to-neuron models on the natural scene experiment and used those models to predict spontaneous activity. **Figure 6.4** shows the same comparison as **Figure 6.3**, but now in mouse V1. In all neurons (left plot), many show the same pattern as in human V1, with equal prediction accuracy across experiments. However, there is a large population of neurons that show much higher prediction accuracy for natural scenes than spontaneous activity. When this is broken down into excitatory (middle) and inhibitory (right) cells and interesting pattern emerges. The majority of neurons that show a prediction accuracy advantage to natural scenes data are excitatory neurons, whereas inhibitory neurons tend to be on par across both experimental conditions.

This is an intriguing and exciting finding that deserves a more directed line of research and experiments. It is difficult to draw strong conclusions from this dataset and inhibitory neurons were not the direct target and there are many fewer than excitatory neurons. However, this result is a prime example of how comparing neuron-to-neuron and voxel-to-voxel models on similar experimental paradigms can lead to a rich set of research directions and questions.

Chapter 7: Summary, Limitations, & Future Directions

7.1 Summary

The central finding of this work is that a linear transformation of activity in source voxels (the voxel-to-voxel model) predicted activity in nearly every target voxel more accurately than an optimized, nonlinear transformation of the stimulus (the DCNN-based stimulus-to-voxel encoding model)(**Figures 3.1, 4.1** left). This finding clearly demonstrates that the stimulus-to-voxel model is blind to one or more “hidden" sources of variance that induce strong correlations between the activities of voxels across the visual hierarchy.

These hidden sources of variance must be endogenous (i.e., not entirely stimulus-dependent) because the voxel-to-voxel model did not predict more accurately than the stimulus-to-voxel model when source and target voxels were located in different brains(**Figure 4.1**, right). Importantly, we have shown that the correlations induced by these hidden, endogenous sources of variance are highly structured. Induced correlations are strongest between voxels with adjacent receptive fields (**Figure 4.2,** top), even when source and target voxels are hierarchically distant (**Figure 4.3**, bottom) and when stimulus-related signal is removed (**Figure 4.9**). Mean activity correlations (**Figure 4.7**) and predictions from Air or White Matter voxels(**Figure 4.8**) cannot account for the increase in prediction accuracy of voxel-to-voxel models.

We then applied our modeling approach to feature map activation values from two DCNN variants. Both the Cifar-10 based network and AlexNet produced patterns of prediction accuracy between layers that differed from the brain (**Figure 5.1)**. This is an indication that the architecture of DCNN-based models needs to be improved upon to fully model brain activity. Next, we extended our approach even further with neuron-to-neuron modeling of mouse V1 neural activity. These exciting results replicated patterns found in lateral voxel-to-voxel models (**Figures 6.1-6.4)**. Finally, we presented preliminary findings connecting unexplained variance to spontaneous intrinsic activity in both human and mouse data (**Figures 6.5 & 6.6**).

7.2 Limitations

One potential source of correlated activity we were unable to account for are eye movements. Although subjects were required to maintain fixation on a central dot during the experiment, small involuntary eye movements might effectively translate the stimulus in a random direction on each trial. These random translations could induce endogenous, spatially correlated and even retinotopically mapped variations in activity that would not be captured by a stimulus-to-voxel model (unless the model somehow incorporated recorded eye movements on each trial; unfortunately, eye movement data is not available for this experiment). This eye-movement-induced variation in activity would most likely be largest in brain areas or regions with small receptive fields and high spatial frequency preference. This would explain why voxel-to-voxel models offered a dramatic improvement in prediction accuracy over the stimulus-to-voxel model for voxels in the foveal representation, and were most effective in low-level visual areas Furthermore, eye-movement-induced variation in activity would most likely be smallest in brain areas or regions with large receptive fields and high spatial frequency preference. This would explain why we observed a decrease in voxel-to-voxel model prediction accuracy with ascent of the visual hierarchy.

However, two aspects of our results challenge the eye-movement as being the main source of correlated activity. First, if low-level areas are more influenced by eye-movement than high-level areas, it should be more difficult to predict the activity of target voxels with a feedback model than a feed-forward model, and the difficulty should increase with hierarchical distance below the source voxels. Instead, we observe that the prediction accuracy of feedback models for any source/target pair is almost always greater than for the corresponding feed-forward model and does not depend upon hierarchical distance. The eye-movement interpretation thus contradicts the feed-forward/feedback asymmetries in prediction accuracy that we observed in our data.

A second challenge to the eye movement explanation is the discrepancy between the results for natural scenes vs. retinotopic mapping experiments. Foveal receptive fields are readily estimated from activity evoked by retinotopic mapping stimuli, but not, as we have shown, by natural scenes. Thus, during retinotopic mapping foveal voxels do not seem to be as dominated by stimulus-independent variance as they are during natural scene stimulation. The fixation task is the same for the retinotopic mapping and natural scenes experiments, so the frequency and magnitude of eye movements during the two experiments is unlikely to differ by much. This discrepancy suggests that the correlations exploited by the voxel-to-voxel model may in fact have more to do with the way that natural scenes (as opposed to synthetic stimuli) are processed than with eye movements. Nevertheless, future work should be careful to take eye movements into account when applying similar methods.

7.3 Future Directions

A compelling interpretation of the superior prediction accuracy of voxel-to-voxel relative to stimulus-to-voxel models is that it reflects the well-known predominance of ongoing activity in the visual system (Kriegeskorte, 2015; Van Den Heuvel & Pol, 2010). An extensive body of work has shown ongoing, stimulus-independent activity to be meaningfully structured (Berkes et al., 2011; Van Den Heuvel & Pol, 2010), highly correlated across neurons and regions (Zhang et al., 2014), in register with cortical topography (Arcaro et al., 2015; Heinzle et al., 2011; Kenet et al., 2003), and not dismissible as eye movement or noise (Arcaro et al., 2015). Additionally, our results on cross-subject prediction suggests the encoding model presented here is near the limit at which any model can leverage solely stimulus-based information to explain variance in the fMRI signal.

Interpreted this way, our results establish that for the vast majority of voxels, and therefore most of visual cortex, ongoing activity is the dominant component of activity measured during vision. This underscores the need for brain models with internal dynamics that can generate structured ongoing activity. In order for voxel-to-voxel models to predict more accurately than the stimulus-to-voxel models, at many places in visual cortex activation at any one time cannot be entirely stimulus dependent, but must reflect the interaction of stimulus-dependent signals with internal state. This internal is likely a combination of many sources of activity. Feedback activity related to memory and attention, interoception signals, information regarding body positioning and movement, and even affective states may be continuously broadcast throughout the cortex. Our results add to a growing body of evidence that incoming stimulus signal is integrated into a complex and dynamic system of ongoing activity rather than the main driver of neural processing.

Furthering this point, the success of neuron-to-neuron models shows that ongoing activity can be leveraged to predict neural activity at both fine and coarse spatial scales. Our flexible modeling approach allows us to connect findings across species and potentially shed light on the link between neuron level activity and signal recorded in fMRI. Additionally, preliminary results in both resting state fMRI and spontaneous activity recorded with 2-Photon imaging pave the way for potential future experiments to reveal how incoming sensory information is integrated into the dynamic intrinsic activity of the brain to seamlessly create our perceptual experience.

References

Andersson, J. L. R., Jenkinson, M., & Smith, S. M. (2010). *Non-linear registration, aka spatial normalisation.* University of Oxford. https://www.fmrib.ox.ac.uk/datasets/techrep/tr07ja2/tr07ja2.pdf

Arcaro, M. J., Honey, C. J., Mruczek, R. E., Kastner, S., & Hasson, U. (2015). Widespread correlation patterns of fMRI signal across visual cortex reflect eccentricity organization. *Elife*, *4*, e03952.

Arieli, A., Shoham, D., Hildesheim, R., & Grinvald, A. (1995). Coherent spatiotemporal patterns of ongoing activity revealed by real-time optical imaging coupled with single-unit recording in the cat visual cortex. *Journal of Neurophysiology*, *73*(5), 2072–2093. https://doi.org/10.1152/jn.1995.73.5.2072

Berkes, P., Orbán, G., Lengyel, M., & Fiser, J. (2011). Spontaneous Cortical Activity Reveals Hallmarks of an Optimal Internal Model of the Environment. *Science*, *331*(6013), 83–87. https://doi.org/10.1126/science.1195870

Biswal, B., Zerrin Yetkin, F., Haughton, V. M., & Hyde, J. S. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magnetic Resonance in Medicine*, *34*(4), 537–541.

Cheng, K. (2018). Exploration of human visual cortex using high spatial resolution functional magnetic resonance imaging. *NeuroImage*, *164*, 4–9.

Community, B. O. (2018). *Blender—A 3D modelling and rendering package*. Blender Foundation. http://www.blender.org

Dale, A., Fischl, B., & Sereno, M. I. (1999). Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction. *NeuroImage*, *9*(2), 179–194.

David, S. V., & Gallant, J. L. (2005). Predicting neuronal responses during natural vision. *Network: Computation in Neural Systems*, *16*(2–3), 239–260.

Dumoulin, S. O., & Wandell, B. A. (2008). Population receptive field estimates in human visual cortex. *Neuroimage*, *39*(2), 647–660.

Friston, K. J. (2011). Functional and effective connectivity: A review. *Brain Connectivity*, *1*(1), 13–36.

Gaglianese, A., Vansteensel, M. J., Harvey, B. M., Dumoulin, S. O., Petridou, N., & Ramsey, N. F. (2017). Correspondence between fMRI and electrophysiology during visual motion processing in human MT+. *NeuroImage*, *155*, 480–489.

Gao, J. S., Huth, A. G., Lescroart, M. D., & Gallant, J. L. (2015). Pycortex: An interactive surface visualizer for fMRI. *Frontiers in Neuroinformatics*, *9*, 23. https://doi.org/10.3389/fninf.2015.00023

Grinvald, A., Slovin, H., & Vanzetta, I. (2000). Non-invasive visualization of cortical columns by fMRI. *Nature Neuroscience*, *3*(2), 105.

Güçlü, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014.

Heinzle, J., Kahnt, T., & Haynes, J.-D. (2011). Topographically specific functional connectivity between visual field maps in the human brain. *Neuroimage*, *56*(3), 1426–1436.

Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., & Smith, S. M. (2012). FSL. *NeuroImage*, *62*(2), 782–790. https://doi.org/10.1016/j.neuroimage.2011.09.015

Jenkinson, M., & Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, *5*(2), 143–156.

Kay, K. N., Naselaris, T., & Gallant, J. L. (2011). *FMRI of human visual areas in response to natural images.* CRCNS.org.

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, *452*(7185), 352.

Kay, K. N., Winawer, J., Mezer, A., & Wandell, B. A. (2013). Compressive spatial summation in human visual cortex. *Journal of Neurophysiology*, *110*(2), 481–494.

Kenet, T., Bibitchkov, D., Tsodyks, M., Grinvald, A., & Arieli, A. (2003). Spontaneously emerging cortical representations of visual attributes. *Nature*, *425*(6961), 954.

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*, 417–446.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 1097–1105.

Logothetis, N. K. (2007). The ins and outs of fMRI signals. *Nature Neuroscience*, *10*(10), 1230–1232. https://doi.org/10.1038/nn1007-1230

Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. *Nature*, *453*(7197), 869–878. https://doi.org/10.1038/nature06976

Nishida, S., Nakano, Y., Blanc, A., & Nishimoto, S. (2019). Brain-mediated Transfer Learning of Convolutional Neural Networks. *CoRR*, *abs/1905.10037*. http://arxiv.org/abs/1905.10037

Pedregosa, F., Eickenberg, M., Ciuciu, P., Thirion, B., & Gramfort, A. (2015). Data-driven HRF estimation for encoding and decoding models. *NeuroImage*, *104*, 209–220.

Raichle, M. E., & Mintun, M. A. (2006). Brain work and brain imaging. *Annual Review of Neuroscience*, *29*(1), 449–476. https://doi.org/10.1146/annurev.neuro.29.051605.112819

Savitzky, Abraham., & Golay, M. J. E. (1964). Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, *36*(8), 1627–1639. https://doi.org/10.1021/ac60214a047

Singh, G., Memoli, F., Ishkhanov, T., Sapiro, G., Carlsson, G., & Ringach, D. L. (2008). Topological analysis of population activity in visual cortex. *Journal of Vision*, *8*(8), 11–11. https://doi.org/10.1167/8.8.11

Strappini, F., Wilf, M., Karp, O., Goldberg, H., Harel, M., Furman-Haran, E., Golan, T., & Malach, R. (2019). Resting-state activity in high-order visual areas as a window into natural human brain activations. *Cerebral Cortex*, *29*(9), 3618–3635.

Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M., & Harris, K. D. (2019a). *Recordings of ten thousand neurons in visual cortex in response to 2,800 natural images* [Data Repository].

https://figshare.com/articles/Recordings_of_ten_thousand_neurons_in_visual_cortex_in_response_to_2_800_natural_images/6845348

Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M., & Harris, K. D. (2019b). High-dimensional geometry of population responses in visual cortex. *Nature*, *571*(7765), 361–365. https://doi.org/10.1038/s41586-019-1346-5

St-Yves, G., & Naselaris, T. (2018). The feature-weighted receptive field: An interpretable encoding model for complex feature spaces. *NeuroImage*, *180*, 188–202.

Trenholm, S., & Krishnaswamy, A. (2020). An Annotated Journey through Modern Visual Neuroscience. *Journal of Neuroscience*, *40*(1), 44–53. https://doi.org/10.1523/JNEUROSCI.1061-19.2019

Tsodyks, M., Kenet, T., Grinvald, A., & Arieli, A. (1999). Linking spontaneous activity of single cortical neurons and the underlying functional architecture. *Science (New York, N.Y.)*, *286*(5446), 1943–1946. https://doi.org/10.1126/science.286.5446.1943

Van Den Heuvel, M. P., & Pol, H. E. H. (2010). Exploring the brain network: A review on resting-state fMRI functional connectivity. *European Neuropsychopharmacology*, *20*(8), 519–534.

Woolrich, M. W., Ripley, B. D., Brady, M., & Smith, S. M. (2001). Temporal Autocorrelation in Univariate Linear Modeling of FMRI Data. *NeuroImage*, *14*(6), 1370–1386. https://doi.org/10.1006/nimg.2001.0931

Wu, M. C.-K., David, S. V., & Gallant, J. L. (2006). Complete functional characterization of sensory neurons by system identification. *Annu. Rev. Neurosci.*, *29*, 477–505.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.

Zhang, D., Wen, X., Liang, B., Liu, B., Liu, M., & Huang, R. (2014). Neural Associations of the Early Retinotopic Cortex with the Lateral Occipital Complex during Visual Perception. *PloS One*, *9*(9), e108557.