# Neural System Identification with Neural Information Flow

Katja Seeliger* Luca Ambrogioni* Yağmur Güçlütürk Umut Güçlü
Marcel A. J. van Gerven

**Abstract**

Neural information flow (NIF) is a new framework for system identification in neuroscience. NIF subsumes population receptive field estimation, neural encoding, effective connectivity analysis and hemodynamic response estimation in a single differentiable model that can be trained end-to-end via stochastic gradient descent. NIF models represent neural information processing systems as a network of coupled tensors, each encoding the representation of the sensory input contained in a brain region. The elements of these tensors can be interpreted as cortical columns whose activity encodes the presence of a specific feature in a spatio-temporal location. Each tensor is coupled to the measured data specific to a brain region via low-rank observation models that can be decomposed into the spatial, temporal and feature receptive fields of a localized neuronal population. Both these observation models and the convolutional weights defining the information processing within regions and effective connectivity between regions are learned end-to-end by predicting the neural signal during sensory stimulation. We trained a NIF model on the activity of early visual areas using a large-scale fMRI dataset. We show that we can recover plausible visual representations and population receptive fields that are consistent with the existing literature. Trained NIF models are accessible for *in silico* analyses.

## 1 Introduction

Uncovering the neural computations that subserve cognition and behaviour is a major goal in neuroscience (Churchland and Sejnowski, 1992). Arguably, a true understanding of the brain requires replicating biological neural information processing *in silico*. However, a general approach for estimating large-scale brain models from observed data has not been proposed so far. This paper introduces a new framework, referred to as *neural information flow* (NIF), which allows us to achieve this goal.

A popular approach for modeling neural information processing is a goal-driven approach, where a basis set of of stimulus features optimized to solve a specific task is used to model neural responses to complex naturalistic input (Naselaris et al., 2011; van Gerven, 2017; Yamins and DiCarlo, 2016). Using this approach, the best results so far have been obtained using deep neural networks (DNNs) (Kriegeskorte, 2015; Güçlü and van Gerven, 2015; Yamins and DiCarlo, 2016; Agrawal et al., 2014; Cichy et al., 2016; Horikawa and Kamitani, 2017; Cadena et al., 2019). However, these models have been optimized on tasks such as object classification on image databases, and not for explaining brain responses. While there exists a correspondence between DNNs and brains at a functional level, they do not provide realistic models of neural information processing.

An alternative data-driven approach is to directly estimate neural models from measurements of neural activity. We refer to this uncovering of neural information processing systems from observed data as *neural system identification* (Stanley, 2005; Wu et al., 2006). This approach has been used to reveal various mechanisms of neural information processing in biological systems (Joukes et al., 2014; Klindt et al., 2017; Antolík et al., 2016; McIntosh et al., 2016; Brackbill et al., 2017). However, so far,

*Equal contribution

neural system identification has been used to explain neural responses in individual brain regions. Instead, we aim to estimate whole-brain models that model neural computations in individual neural populations as well as causal interactions between neural populations.

Both goal-driven and data-driven models tend to ignore the causal interactions between brain regions. That is, they are lacking the coupling between distinct neural regions commonly studied with effective connectivity methods (Friston et al., 2003; Friston, 2011; Liao et al., 2010; Ambrogioni et al., 2017). Established techniques for uncovering effective connectivity are able to uncover this causal coupling (Friston et al., 2003), but do not capture the nature of information processing that drives the interaction between brain regions.

NIF models can be interpreted as "synthetic (*in silico*) brain models" that learn to capture the nonlinear computations that take place in real brains. Instead of modeling brain regions in an isolated fashion, they include connectivity between brain regions by taking afferent input into account (Haak et al., 2013). Moreover, by making use of convolution and factorization, we are able to estimate whole-brain models in an efficient manner. In the following, we outline the basic methodology of NIF modeling. Using a large functional magnetic resonance imaging (fMRI) dataset acquired under naturalistic stimulation we demonstrate that the model is capable of generating realistic brain measurements and that the computations captured in the model are biologically meaningful.

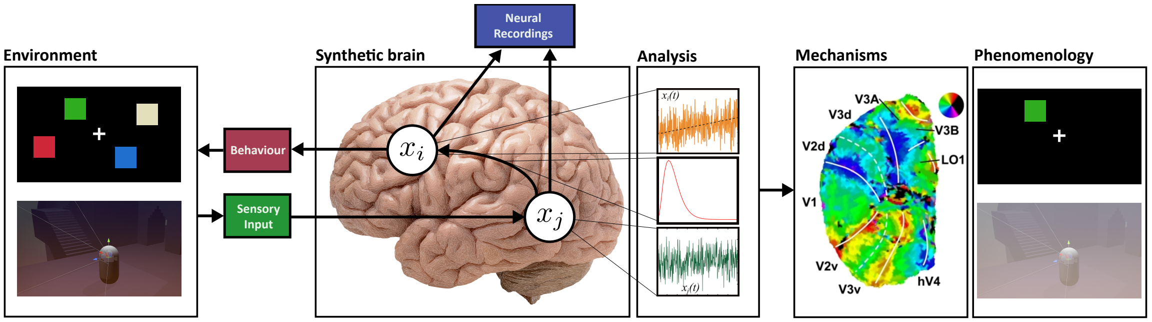

Figure 1: The philosophy underlying neural information flow. NIF models define synthetic brains that model information processing in real brains. They are specified in terms of mutually interacting neuronal populations (white discs) that receive sensory input (green) and give rise to measurements of neural activity (blue) and/or behavior (red). In practice, NIF models may consist of up to hundreds such interacting populations. They can be estimated by fitting them to neurobehavioural data acquired under these tasks. By analyzing NIF models, we can gain a mechanistic understanding of neural information processing in real brains and how neural information processing relates to phenomenology.

This paper outlines the basic principles of NIF models. However, the framework is general in the sense that the experimenter is free to choose the neural architecture of individual brain regions and how these regions map onto observed measurements, which can be either neural or behavioural in nature. The philosophy of neural information flow is outlined in Figure 1. Given the generality of our framework, we expect that it will guide the development of a new family of generative models that allow us to uncover the principles of neural computations in biological systems.

## 2 Neural information flow

The purpose of a NIF model is to capture the neuronal computations that take place within and between neuronal populations in response to a naturalistic sensory input. The core of a NIF model is a deep convolutional architecture. Each layer of this architecture stores the representation of the sensory input encoded in a specific brain region. Information is processed through convolutions, which model the topographically organized connectivity between brain regions. These representations are used to predict measurements through low-rank observation models that couple each layer of the network to the observable responses of a specific brain region. These low-rank observation models can be factorized into spatial, temporal and feature components.

The model parameters are estimated by fitting the measured neural signals during sensory stimulation. Specifically, the model receives the same sensory input that is presented to the participant and predicts the measurements of all the brain regions of interest. Both the internal convolutions and the observation models are trained end-to-end using SGD. Note that the only error signal comes from the neural responses without any further pre-training and regularization. This differs from most existing encoding approaches, where neural responses are predicted from the activations of a network trained on an unrelated classification task (Kriegeskorte, 2015; Güçlü and van Gerven, 2015; Agrawal et al., 2014; Horikawa and Kamitani, 2017). In the following we describe the components in more detail.

## 2.1 Modeling neural representations

We model the neural representations encoded in individual brain regions using tensors. The activity of a neural area is encoded in a four-dimensional tensor $\mathbf{N} \in \mathbb{R}^{N_c \times N_x \times N_y \times N_t}$, whose array dimensions represent channels $c$, two spatial coordinates $(x, y)$ and time $t$, respectively. During training on neural measurements, the feature maps $\mathbf{N}[i, :, :, :]$ learn to encode neural processing of specific stimulus (input) characteristics such as oriented edges or coherent motion. Consequently, a tensor element can be interpreted as the response of one cortical column. Under the same interpretation, cortical hyper-columns are represented by a sub-tensor storing the activations of all the columns that respond to the same spatial location. Sensory inputs are represented in the same manner. The input tensor represents the responses of sensory receptors such as the photoreceptors in the retina.

## 2.2 Modeling information flow

We model effective connectivity from a source region $a$ to a target region $b$ as a convolution between the neural tensor $\mathbf{N}_a$ and a tensor of synaptic weights $\mathbf{W}_{a\to b}$:

$$(\mathbf{N}_a \star \mathbf{W}_{a\to b})[c_{\text{out}}, x, y, t] = \sum_{c_{\text{in}}, dx, dy, dt} \mathbf{N}_a[c_{\text{in}}, x - dx, y - dy, t - dt]\mathbf{W}_{a\to b}[c_{\text{in}}, dx, dy, dt, c_{\text{out}}] \ . \quad (1)$$

In other words, here we model neural processing using 3D convolutions. To enforce causality of the neural responses, the temporal filters should be causal, meaning that the only non-zero weights correspond to past time points. However, this assumption can be dropped when the time scale of our observations is much slower than that of the underlying temporal dynamics.

As shown in Equation (1), the tensor $\mathbf{W}$ has five array dimensions: input channels, two spatial coordinates, one temporal coordinate and one output channel. The convolution is performed along the two spatial dimensions and the temporal dimension. The spatial weights model the topographically organized synaptic connections while the temporal component models synaptic delays. Using this notation, we can define the activation of the $j$-th brain area as a function of its afferent input:

$$\mathbf{N}_j = \mathbf{f}\left(\sum_{i=1}^{N} \mathbf{N}_i \star \mathbf{W}_{i\to j} + \mathbf{B}_j\right), \quad (2)$$

where $\mathbf{f}$ is a (nonlinear) activation function (applied element-wise) and $\mathbf{B}_j$ determines the bias. Using this setup, we can model how neural populations respond to sensory input, as well as to each other. Note further that bottom-up and top-down interactions between brain regions can be integrated in the same model.

### 2.3 Modeling observable signals

NIF models are estimated by linking neural tensors to observation models that capture (indirect) measurements of brain activity. Observations are represented using tensors $\mathbf{Y}$ that reflect multiple responses in space and time. They are modelled as

$$\mathbf{Y} = \mathbf{g}\left(\mathbf{N}_1, \ldots, \mathbf{N}_N\right) + \boldsymbol{\epsilon} \, , \tag{3}$$

where $\mathbf{g}$ is the observation model and $\boldsymbol{\epsilon}$ is measurement noise. The exact form of $\mathbf{g}$ depends on the kinds of measurements that are being made. Neuroimaging methods such as fMRI, single- and multi-unit recordings, local field potentials, calcium imaging, EEG, MEG; but also motor responses and eye movements are observable responses to afferent input and can thus be used as a training signal. Note that the same brain regions can be observed using multiple observation models, conditioning them on multiple heterogeneous datasets at the same time. This provides a solution for multimodal data fusion in neuroscience (Uludağ and Roebroeck, 2014).

In this paper, we focus on blood-oxygenation-level dependent (BOLD) responses obtained for individual voxels using fMRI. In this case, we can consider the voxel responses separately for each region, such that we have $\mathbf{Y}_i = \mathbf{g}_i\left(\mathbf{N}_i\right) + \boldsymbol{\epsilon}$ for each region $i$. Let $\mathbf{Y}_i \in \mathbb{R}^{K \times T}$ denote BOLD responses of $K$ voxels acquired over $T$ time points inside the $i$-th region. Our observation model for the $k$-th voxel in that region is defined as

$$\mathbf{Y}_i[k, t + \Delta_t] = \sum_{c,x,y,t} \mathbf{N}_i[c, x, y, t]\mathbf{U}_i[c, x, y, t, k] + \epsilon[k] \, , \tag{4}$$

where $\Delta_t$ is a temporal shift of the BOLD response that is used to take a basic offset in the hemodynamic delay into account (4.9 s in our example). Every brain region can be observed using a function of the form shown in Equation (4).

To simplify parameter estimation and facilitate model interpretability we use a factorized representation of $\mathbf{U}$. That is,

$$\mathbf{U}[c, x, y, t, k] \approx \mathbf{U}_c[c, k]\mathbf{U}_s[x, y, k]\mathbf{U}_t[t, k] \, , \tag{5}$$

where $\mathbf{U}_c[\cdot, k]$ are the channel loadings which capture the sensitivity of a voxel to specific input features, $\mathbf{U}_s[\cdot, \cdot, k]$ is the spatial receptive field of a voxel and $\mathbf{U}_t[\cdot, k]$ is the temporal profile of the observed BOLD response of the $k$-th voxel. Hence, the estimated voxel-specific observation models have a direct biophysical interpretation.

We further facilitate parameter estimation by using a spatial weighted low-rank decomposition of the spatial receptive field:

$$\mathbf{U}_s[x, y, k] \approx b_k + \sum_{r=1}^{R} a_{k,r}\mathbf{U}_{x,r}[x, k]\mathbf{U}_{y,r}[y, k] \; . \tag{6}$$

Here, $b_k$ is a voxel-specific bias and $a_{k,r}$ are positive rank amplitudes. In our experiments, we used $R = 4$. To further stabilize the model and obtain localized population receptive fields, we apply a softmax nonlinearity to the columns (voxel-specific weights) of $\mathbf{U}_t$, $\mathbf{U}_x$ and $\mathbf{U}_y$. That is, the elements $u_i$ of each column vector $\mathbf{u}$ of these matrices are given by

$$u_i = \sigma_i(\mathbf{v}) = \exp(v_i) / \sum_j \exp(v_j) \; , \tag{7}$$

where the $v_i$ are learnable parameters. This enforces positively weighted spatio-temporal receptive fields and reduces noise in the individual estimation of the voxel-specific weight values. The rank limits the complexity of the spatial observation model. Rank one models can estimate unimodal receptive fields. However, a small number of voxels have nonclassical receptive fields that respond to multiple parts of the input space (see Figure 6), for which more degrees of freedom are needed.

### 2.4 Model estimation

Once the architecture of the NIF model is defined, its parameters (synaptic weights and observation model parameters) can be estimated using stochastic gradient descent (SGD). NIF allows modeling recurrent neural computations (loops) that arise due to both bottom-up and top-down connections between brain regions as well by unrolling the graph and performing backpropagation through time (Werbos, 1990). However, in the present paper we only model feed-forward processing. Losses are estimated within brain regions as mean squared error across voxels. The individual loss terms for every brain region are summed to obtain a final loss that is minimized using SGD. Note that since the model couples neuronal populations, region-specific estimates are constrained by one another and consequently make use of all observed data. The NIF example presented here was implemented in the `chainer` framework for automatic differentiation (Tokui et al., 2015).

## 3 Experimental validation

To test our example visual system model, we made use of a unique large-scale functional MRI dataset for which which one subject was exposed to almost 23 hours of complex naturalistic spatiotemporal stimuli. Specifically, we presented episodes from the BBC series *Doctor Who* (Davies et al., 2005).

### 3.1 Stimulus material

A single human participant (male, age 27.5) watched 30 episodes from seasons 2 to 4 of the 2005 relaunch of *Doctor Who*. This comprised the training set which was used for model estimation. Episodes were split into 12 min chunks (with each last one having varying length) and presented with a short break after every two runs. The participant additionally watched repeated presentations of short movies (*Pond Life* (five movies of 1 min, 26 repetitions) and *Space / Time* (two movies of 3 min, 22 repetitions)) in random permutations after nearly every episode. They were taken from the series' sequel to avoid overlap with the training data. This comprised the test set which was used for model validation.

### 3.2 Data acquisition

We collected 3T whole-brain fMRI data. It was made sure that the training stimulus material was novel to the participant. Data were collected inside a Siemens 3T MAGNETOM Prisma system using a 32-channel head coil (Siemens, Erlangen, Germany). A T2*-weighted echo planar imaging pulse sequence was used for rapid data acquisition of whole-brain volumes (64 transversal slices with a voxel size of $2.4 \times 2.4 \times 2.4$ mm$^3$ collected using a TR of 700 ms). We used a multiband-multi-echo protocol with multiband acceleration factor of 8, TE of 39 ms and a flip angle of 75 degrees. The video episodes were presented on a rear-projection screen with the Presentation software package, cropped to $698 \times 732$ pixels squares so that they covered 20° of the vertical and horizontal visual field. The participant's head position was stabilized within and across sessions by using a custom-made MRI-compatible headcast, along with further measures such as extensive scanner training. The participant had to fixate on a fixation cross in the center of the video. At the beginning of every break and after every test set video a black screen was shown for 16 s to record BOLD fadeout, however here we omit this part of the data. In total this leaves us with 118.197 whole-brain volumes of single-presentation data, forming our training set (used for model estimation); and 1.031 volumes of resampled data, forming our test set (used for model evaluation).

Data collection was approved by the local ethical review board. All specifics of the data set are described in a separate manuscript accompanying the data that will be made publicly available.

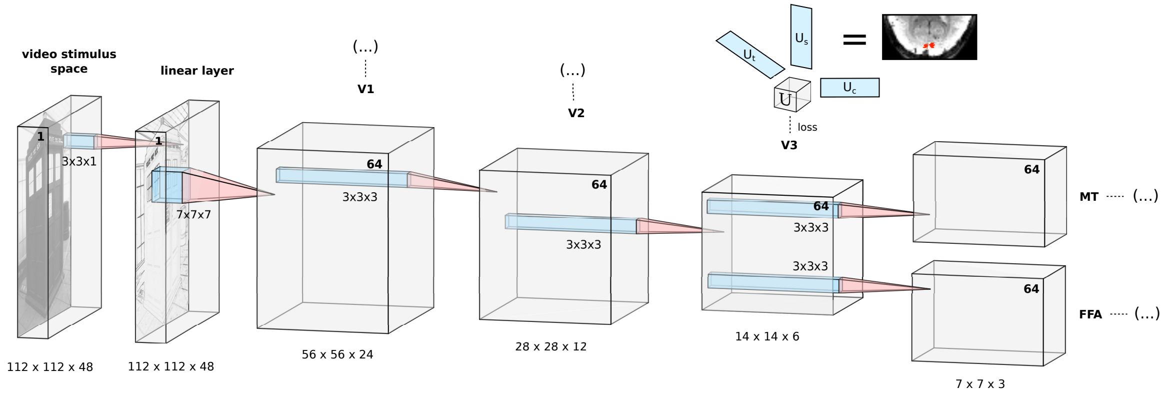

Figure 2: The example NIF model represents a simplified feed-forward architecture of early visual system regions. Underneath the tensors resulting from the 3D convolution operations we state the size of each input space ($x \times y \times t$) to the next layer. The number of feature maps (channels) in each input space is printed in boldface, with the stimulus (input) space consisting of a single channel. The input to the network are 3D stimulus video segments consisting of $3 \times 16$ frames (covering three TRs of 700 ms each), aligned with the hemodynamic response by applying a fixed delay of 7 TRs. The first convolutional layer is not attached to a region observation model, but is a single-channel linear spatial convolution layer. It serves as a learnable linear preprocessing step that accounts for retinal and LGN modulations. Convolutional kernel sizes are $7 \times 7 \times 7$ in the second convolutional layer (leading to the V1 tensor), and $3 \times 3 \times 3$ for all other layers. After every convolution operation we apply a sigmoid nonlinearity and spatial average pooling with $2 \times 2 \times 2$ kernels. Before entering the $\mathbf{U}_t$ observation models the temporal dimension is average pooled so that each point $t$ covers one TR. All weights in this model (colored blue) are learned by backpropagating the mean squared error losses from predicting the BOLD activity of the observed voxels. The voxel-specific observation models consisting of the spatio-temporal weight vectors $\mathbf{U}_s$ and $\mathbf{U}_t$ and the channel observation vector $\mathbf{U}_c$ enable the end-to-end training of the model from observational data.

## 3.3 Data preprocessing

Minimal BOLD data preprocessing was performed using `FSL v5.0`. Volumes were first aligned within each 12 min run to their center volume (run-specific reference volume). Next, all reference volumes were aligned to the center volume of the first run (global reference volume). The run-specific transformations were applied to all volumes to align them with the global reference volume.

The signal of every voxel used in the model was linearly detrended, then standardized (demeaning, unit variance) per run. Test set BOLD data was averaged over repetitions to increase signal to noise ratio, and as a final step the result was standardized again. A fixed delay of 7 TRs (4.9 s) was used to associate stimulus video chunks with responses and allow the model to learn voxel-specific HRF delays within $\mathbf{U}_t$. With the video segments covering 3 TRs starting from the fixed delay, the BOLD signal corresponding to a stimulus is thus expected to occur within a time window of 4.9 s to 6.3 s after the onset of the segment. As there were small differences between frame rates in the train and test sets we converted the stimulus videos to a uniform frame rate of 23.86 Hz (16 frames per TR) for training the example model. To reduce model complexity we downsampled the videos to $112 \times 112$. As the model operates on three consecutive TRs, the training input size was $112 \times 112 \times 48$. The stimuli were converted to greyscale prior to presenting them to the model. Otherwise stimuli were left just as they were presented in the experiment.

## 3.4 Model architecture

We implemented a purely feed-forward architecture for modeling parts of the visual system (LGN, V1, V2, V3, FFA and MT). The used architecture is illustrated in detail in Figure 2. FFA and MT have their own tensors originating from V3 to allow for a simplified model of the interactions between upstream and downstream areas. We intentionally used a simplified model to focus on demonstrating the capabilities of the NIF framework. To model LGN output, we used a linear layer consisting of a single $3 \times 3 \times 1$ spatial convolutional kernel. The model was trained for 11 epochs with a batch size of 3, using the Adam optimizer (Kingma and Ba, 2014) with learning rate $\alpha = 5 \times 10^{-4}$. Weights were initialized with Gaussian distributions scaled by the number of feature maps in every layer (He et al., 2015).

# 4 Results

In this paper, we focus on the processing of visual information. In the following, we show that a NIF model uncovers meaningful characteristics of the visual system.

## 4.1 Response prediction

After training the NIF model, we observed that BOLD responses in a majority of voxels in each brain region could be significantly predicted by the model ($p < 0.01$, Bonferroni-corrected for the total number of gray matter voxels). This is illustrated in Figure 3, showing voxel-wise correlations between predicted and observed test data per region. The results show that the NIF model indeed generates realistic brain activity in response to unseen input stimuli (out-of-sample prediction).

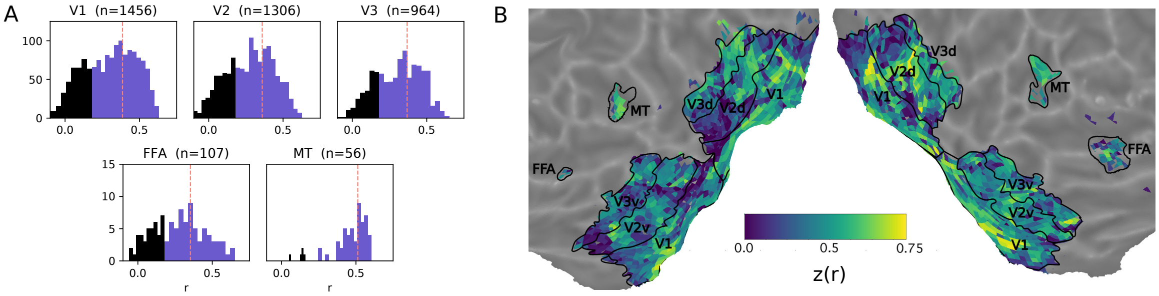

Figure 3: A. Histograms of voxel-wise correlations between predicted and observed BOLD responses on the test set in different observed brain regions. The vertical line marks the median. The blue area shows the significantly predictable voxels. B. Cortical flatmap of the distribution of all correlations across the visual system. For the map we applied a Fisher z-transform to facilitate linear visual comparison of correlation magnitudes.

### 4.1.1 Linear preprocessing

Both the retina and the LGN process visual information before it enters the visual cortex (Graham et al., 2006; Dan et al., 1996). In our NIF model, we account for this biological preprocessing with a learnable linear purely spatial single-channel convolutional layer ($3 \times 3 \times 1$) (without nonlinearities) that connects the retinal input to the V1 input. In the main experiment presented in this paper, we found that this layer behaves as an edge extractor (Figure 4B).

In alternative training runs, we included a second preprocessing channel in order to investigate its effect. In this case, one channel became an edge extractor while the other channel extracted luminance (Figure 4C). This is likely to be a reflection of the independence of luminance and contrast information in natural images and in LGN responses (Mante et al., 2005). These results indicate that the model is capable of learning a biologically plausible linear transformation of the visual input.

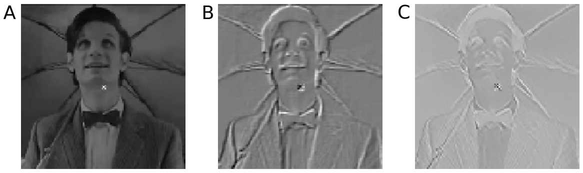

Figure 4: Learned linear preprocessing. A. Original stimulus. B. Using one channel, the model learns to extract edges. C. When an additional channel is added, the model learns to capture luminance information.

To simplify the model and reduce computational burden we used a single preprocessing channel in the following analyses.

## 4.2 Neural representations

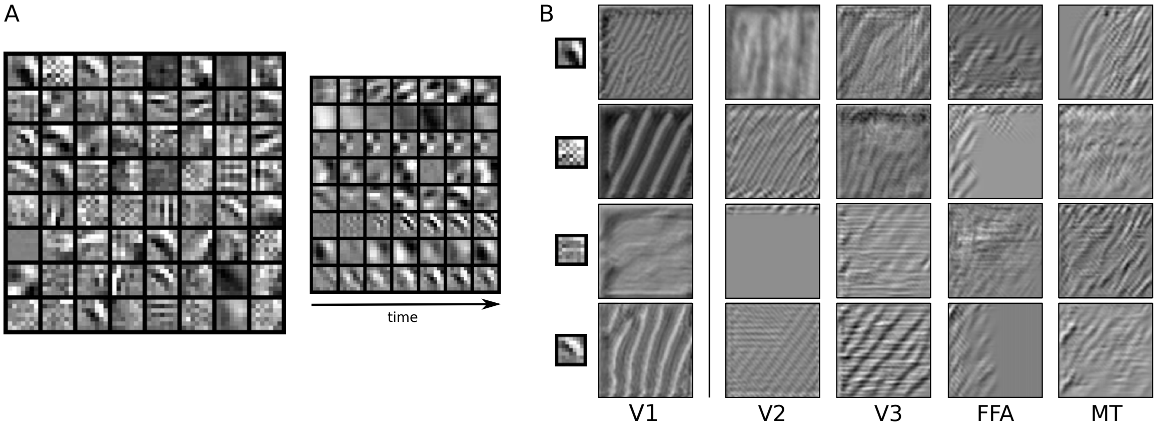

Figure 5: A. The 64 spatio-temporal channel weights estimated from neural data for area V1 (frame three out of seven). The right panel shows seven of these feature weights over time. For visualization, channel weights were clipped at the extremes and all weights were globally rescaled between 0 and 1. B. Preferred inputs maximizing the responses of the first four feature maps for V1 (together with their corresponding channel visualization) and for the feature maps associated with downstream brain regions.

Before nonlinearities are applied, neural network features can be inspected by visualizing the learned weights. Figure 5A shows the 64 channels (feature detectors) learned by the neural tensor connected to V1 voxels. We see that several well-known feature detection mechanisms of V1 arise such as Gabor-like response profiles (Jones and Palmer, 1987). Several of these feature detectors show distinct dynamic temporal profiles (see Figure 5A, right panel), reflecting the processing of visual motion (Joukes et al., 2014).

For higher-order regions we need to resort to other visualization techniques. One common approach is computing the gradient that leads to an increase in the activity in the whole feature map of the target channel, and using this gradient to modify the input, starting from a white noise pattern. The resulting visualizations give an impression of the inputs which specific neural network channels prefer or respond strongly to. This procedure leads to noisy visualizations, but can be stabilized by iteratively repeating it on multiple increasing scales of the input image (Mordvintsev et al., 2015). We have used this procedure to synthesize preferred inputs of our network. We have used four scales (starting at 37 × 37 × 8 increasing with a scaling factor of 1.3 up to 81 × 81 × 18), using the mean absolute error on the feature map activations as loss, and an SGD optimizer with a learning rate of 2000 and L2 regularization. The results for different areas can be seen in Figure 5B. The V1 preferred input synthesis mainly shows similarly oriented bars across the feature maps. Higher-order channels of the model show complex combinations consisting of multiple spatial frequencies.

### 4.3 Receptive field mapping

We examined whether the retinotopical organization of the visual cortex can be recovered from the spatial observation models.

$\mathbf{U}_s$ represents spatial receptive field estimates for every voxel. Some of these voxel-specific receptive fields are shown in Figure 6A. The model has primarily learned classical local unimodal population receptive fields, but has also learned more complex non-classical response profiles (Olshausen and Field, 2005). This matches the expectation that a population (voxel) response is not necessarily restricted to unipolar receptive fields.

To check that the NIF model has indeed captured sensible retinotopic properties we determined the center of mass of the spatial receptive fields, and transformed these centers to polar coordinates using the central fixation point as origin. Sizes of the receptive fields were estimated as the standard deviation across $\mathbf{U}_s$, using the centers of mass as mean. Voxels whose responses could not be significantly predicted were excluded from this analysis.

Figure 6 shows polar angle (B), eccentricity (C) and receptive field size (D) for early visual system areas observed by our model. Maps were generated with `pycortex` (Gao et al., 2015). Note that the boundaries between visual areas V1, V2 and V3 have been estimated with data from a classical wedge and ring retinotopy session. It becomes clear that reversal boundaries align well with the traditionally estimated ROI boundaries. The larger eccentricity and increase in receptive field size (C) matches the expected fovea-periphery organization as well.

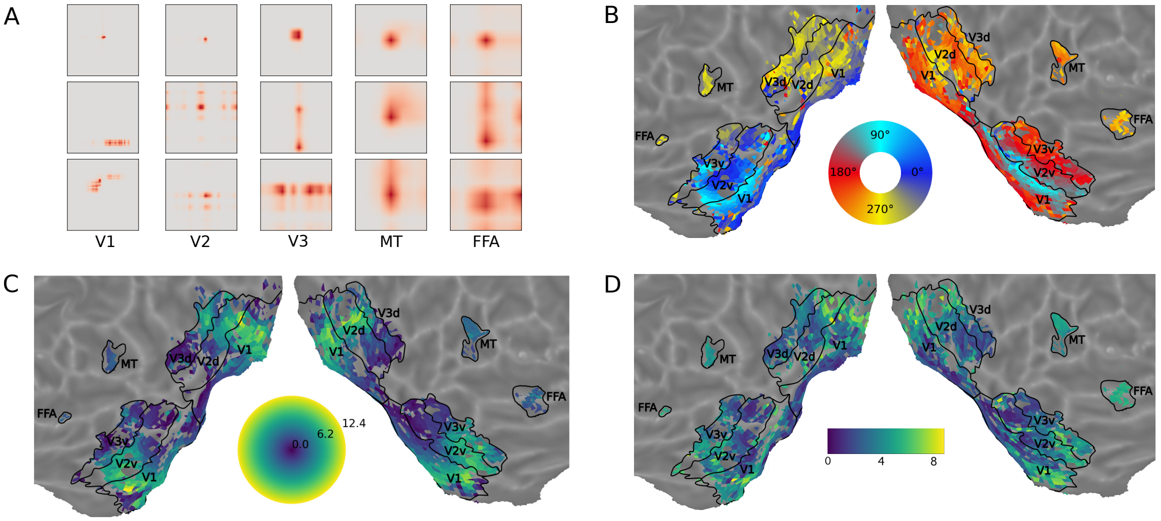

Figure 6: A. Various spatial receptive fields in video pixel space $\mathbf{U}_s$ learned for different ROIs within our framework. Most estimated spatial receptive fields are unipolar. B-D. Basic retinotopy that arose in the voxel-specific spatial observation matrix $\mathbf{U}_s$ within the NIF model. B. Polar angle. C. Eccentricity. D. Receptive field size.

The NIF model thus learns sensible retinotopic characteristics of the visual system directly from naturalistic data. Our results also indicate that the NIF framework allows the estimation of accurate retinotopic maps from naturalistic videos.

# 5 Discussion

This paper introduced neural information flow as a new approach for neural system identification. The approach relies on a neural architecture specified in terms of interacting brain regions that each embody nonlinear computations. By conditioning each brain region on associated measurements of neural activity, we can estimate neural information processing systems end-to-end. By allowing interactions between brain regions, each brain region will act as a regularizer for the neural computations that emerge in other brain regions. After all, the estimated neural computations must jointly explain observed responses across all brain regions.

We showed using fMRI data collected during prolonged naturalistic stimulation that we can successfully predict BOLD responses across different brain regions. Furthermore, meaningful receptive fields emerged after model estimation. Importantly, the learnt receptive fields are specific to each brain region but collectively explain all of the observed measurements. To the best of our knowledge, these results demonstrate for the first time that biologically meaningful information processing systems of multiple interconnected regions can be directly estimated from neural data. NIF allows neuroscientists to specify hypotheses about neuronal interactions and test these by quantifying how well the resulting models explain observed measurements.

NIF generalizes current encoding models. For example, basic population receptive field models (Dumoulin and Wandell, 2008) and more advanced neural network models (van Gerven, 2017) are special cases of NIF that assume no interactions between brain regions and make specific choices for the nonlinear transformations that capture neuronal processing. Furthermore, current approaches mainly rely on using neural networks that are trained to solve another task such as object recognition (Güçlü and van Gerven, 2015; Yamins et al., 2014). An exception is, e.g. (Güçlü and van Gerven, 2016), but the employed models required the transformation of sensory input to lower-dimensional stimulus features and did not allow the explicit recovery of neural computations in individual regions of interest. Here we show that biologically-interpretable models can be estimated directly from neural data using neural information flow.

The present work also provides a new approach to effective connectivity analysis. The researcher can specify alternative NIF models and then use explained variance as a model selection criterion. This is similar in spirit to dynamic causal modelling. However, instead of using changes in neural *dynamics* to estimate effective connectivity, we can embraces changes in neural *computation* to estimate causal interactions.

NIF can be naturally extended in several ways. The employed convolutional layer to model neural computation can be replaced by neural networks that have a more complex architecture. For example, recurrent neural networks can be used to explicitly model the changes in neural dynamics that are now captured by 3D convolutions. Furthermore, lateral and feedback processes are easily added by adding additional links between brain regions.

NIF models can also be extended to handle other data modalities. Alternative observation models can be formulated that allow us to infer neural computations from other measures of neural activity (e.g., single- and multi-unit recordings, local field potentials, calcium imaging, EEG, MEG). Moreover, NIF models can be conditioned on multiple heterogeneous datasets at the same time, providing an elegant solution for multimodal data fusion. Cortical flow can also be easily applied to other sensory inputs. For example, auditory areas can be conditioned on auditory input (see e.g. (Güçlü et al., 2016)). If this is combined with visual input then we may be able to uncover new properties of multimodal integration (Simanova et al., 2014).

Note that we are not restricted to conditioning NIF models on neural data. We may instead (or also) condition these models on behavioural data, such as motor responses or eye movements. The resulting models should then show the same behavioural responses as the system under study. We can even teach NIF models to solve the task at hand directly using reinforcement learning (Sutton and Barto, 2017). In this sense, NIF models provide a starting point for creating brain-inspired AI systems that more closely model how real brains solve cognitive tasks.

Estimated NIF models can be interpreted as synthetic brains that model their biological counterparts. This implies that we can subject them to any approach which can also be used to probe neural information processing in real brains. For instance, we can apply any method for neural data analysis to the neural time series that result from driving the model with external input. Recently developed nonlinear decoding techniques can shed further light on the neural representations that are encoded by different brain regions (Güçlütürk et al., 2017), providing insight into the phenomenological experience of synthetic brains. Here we restricted ourselves to demonstrating the virtues of our approach using basic receptive field analyses.

Finally, we can use NIF models as in-silico models to examine changes in neural computation. For example, we can examine how neural representations change during learning or by virtual lesioning of the network (Graziano and Aflalo, 2007). This can provide insights into cognitive development and decline. We can also test what happens to neural computations when we directly drive individual brain regions with external input. This provides new approaches for understanding how brain stimulation modulates neural information processing, guiding the development of future neurotechnology (Roelfsema et al., 2018).

Summarizing, we view neural information flow as a starting point for building a new family of rich, general, biologically-inspired computational models that capture neural information processing in biological systems. As such it provides a perfect blend of computational and experimental neuroscience (Churchland and Sejnowski, 2016). NIF models are also scalable since they make use of efficient stochastic gradient methods, as developed by the artificial intelligence community. This provides us with a principled approach to make sense of the high-resolution datasets produced by continuing advances in neurotechnology (Stevenson and Kording, 2011). We expect that (variants of) NIF models will provide exciting new insights into the principles and mechanisms that dictate neural information processing in biological systems.

# Acknowledgements

This research was supported by VIDI grant number 639.072.513 of The Netherlands Organization for Scientific Research (NWO).

# References

Agrawal, P., Stansbury, D., Malik, J., and Gallant, J. L. (2014). Pixels to voxels: Modeling visual representation in the human brain. *arXiv preprint arXiv:1407.5104*.

Ambrogioni, L., Hinne, M., van Gerven, M. A. J., and Maris, E. (2017). GP CaKe: Effective brain connectivity with causal kernels. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 950–959.

Antolík, J., Hofer, S. B., Bednar, J. A., and Mrsic-Flogel, T. D. (2016). Model constrained by visual hierarchy improves prediction of neural responses to natural scenes. *PLoS Computational Biology*, 12(6):1–22.

Brackbill, N., Heitman, A., Sher, A., and Litke, A. (2017). Multilayer recurrent network models of primate retinal ganglion cell responses. In *International Conference on Learning Representations (ICLR)*.

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., and Ecker, A. S. (2019). Deep convolutional models improve predictions of macaque V1 responses to natural images. *PLoS Computational Biology*, 15(4):e1006897.

Churchland, P. S. and Sejnowski, T. J. (1992). *The Computational Brain.* MIT Press.

Churchland, P. S. and Sejnowski, T. J. (2016). Blending computational and experimental neuroscience. *Nature Reviews Neuroscience*, 17(11):667–668.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., and Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6(27755).

Dan, Y., Atick, J. J., and Reid, R. C. (1996). Efficient coding of natural scenes in the lateral geniculate nucleus: Experimental test of a computational theory. *Journal of Neuroscience*, 16(10):3351–3362.

Davies, R. T., Gardner, J., Moffat, S., Young, M., and Collinson, P. (2005). Doctor Who.

Dumoulin, S. O. and 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.

Friston, K. J., Harrison, L., and Penny, W. (2003). Dynamic causal modelling. *NeuroImage*, 19(4):1273–1302.

Gao, J. S., Huth, A. G., Lescroart, M. D., and Gallant, J. L. (2015). Pycortex: An interactive surface visualizer for fMRI. *Frontiers in Neuroinformatics*, 9:23.

Graham, D. J., Chandler, D. M., and Field, D. J. (2006). Can the theory of “whitening” explain the center-surround properties of retinal ganglion cell receptive fields? *Vision Research*, 46(18):2901–2913.

Graziano, M. S. A. and Aflalo, T. N. (2007). Mapping behavioral repertoire onto the cortex. *Neuron*, 56:239–251.

Güçlü, U., Thielen, J., Hanke, M., and van Gerven, M. A. J. (2016). Brains on beats. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2101–2109.

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

Güçlü, U. and van Gerven, M. A. J. (2016). Modeling the dynamics of human brain activity with recurrent neural networks. *Frontiers in Computational Neuroscience*, pages 1–19.

Güçlütürk, Y., Güçlü, U., Seeliger, K., Bosch, S., van Lier, R., and van Gerven, M. A. J. (2017). Reconstructing perceived faces from brain activations with deep adversarial neural decoding. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4249–4260.

Haak, K. V., Winawer, J., Harvey, B. M., Renken, R., Dumoulin, S. O., Wandell, B. A., and Cornelissen, F. W. (2013). Connective field modeling. *NeuroImage*, 66:376–384.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of the 2015 IEEE International Conference on Computer Vision*, pages 1026–1034.

Horikawa, T. and Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, 8(15037).

Jones, J. P. and Palmer, L. A. (1987). An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58:1233–1258.

Joukes, J., Hartmann, T. S., and Krekelberg, B. (2014). Motion detection based on recurrent network dynamics. *Frontiers in Systems Neuroscience*, 8:239.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Klindt, D. A., Ecker, A. S., Euler, T., and Bethge, M. (2017). Neural system identification for large populations separating “what” and "where”. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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.

Liao, W., Mantini, D., Zhang, Z., Pan, Z., Ding, J., Gong, Q., Yang, Y., and Chen, H. (2010). Evaluating the effective connectivity of resting state networks using conditional Granger causality. *Biological Cybernetics*, 102(1):57–69.

Mante, V., Frazor, R. A., Bonin, V., Geisler, W. S., and Carandini, M. (2005). Independence of luminance and contrast in natural scenes and in the early visual system. *Nature Neuroscience*, 8(12):1690.

McIntosh, L. T., Maheswaranathan, N., Nayebi, A., Ganguli, S., and Baccus, S. A. (2016). Deep learning models of the retinal response to natural scenes. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Mordvintsev, A., Olah, C., and Tyka, M. (2015). Inceptionism: Going deeper into neural networks. *Google Research Blog*.

Naselaris, T., Kay, K. N., Nishimoto, S., and Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage*, 56(2):400–410.

Olshausen, B. A. and Field, D. J. (2005). How close are we to understanding V1? *Neural Computation*, 17:1–34.

Roelfsema, P. R., Denys, D., and Klink, P. C. (2018). Mind reading and writing: The future of neurotechnology. *Trends in Cognitive Sciences*, 22(7):1–13.

Simanova, I., Hagoort, P., Oostenveld, R., and van Gerven, M. (2014). Modality-independent decoding of semantic information from the human brain. *Cerebral Cortex*, 24:426–434.

Stanley, G. B. (2005). Neural system identification. In *Neural Engineering*, pages 367–388. Springer.

Stevenson, I. H. and Kording, K. P. (2011). How advances in neural recording affect data analysis. *Nature Neuroscience*, 14(2):139–142.

Sutton, R. S. and Barto, A. G. (2017). *Reinforcement Learning: An Introduction*. The MIT Press, Boston, MA.

Tokui, S., Oono, K., Hido, S., and Clayton, J. (2015). Chainer: A next-generation open source framework for deep learning. In *Workshop on Machine Learning Systems (LearningSys) during Advances in Neural Information Processing Systems (NeurIPS)*.

Uludağ, K. and Roebroeck, A. (2014). General overview on the merits of multimodal neuroimaging data fusion. *NeuroImage*, 102:3–10.

van Gerven, M. A. J. (2017). A primer on encoding models in sensory neuroscience. *Journal of Mathematical Psychology*, 76:172–183.

Werbos, P. J. (1990). Backpropagation through time: What it does and how to do it. *Proceedings of the IEEE*, 78:1550–1560.

Wu, M. C.-K., David, S. V., and Gallant, J. L. (2006). Complete functional characterization of sensory neurons by system identification. *Annual Review of Neuroscience*, 29(1):477–505.

Yamins, D. L. K. and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and 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.