# Estimating flexible across-area communication with neurally-constrained RNN

**Joao Barbosa (palerma@gmail.com), Adrian Valente (adrianvalente16@gmail.com)**
Group for Neural Theory, École Normale Supérieure, Paris, France

**Scott Brincat (sbrincat@mit.edu), Earl Miller (ekmiller@mit.edu)**
The Picower Institute for Learning and Memory, Massachusetts Institute of Technology, Cambridge, MA, USA

**Srdjan Ostojic (srdjan.ostojic@ens.fr)**
Group for Neural Theory, École Normale Supérieure, Paris, France

## Abstract

**Neural computations supporting complex behaviors involve multiple brain regions, and large-scale recordings from animals engaged in complex tasks are increasingly common. A current challenge in analysing these data is to identify which part of the information contained within a brain region is shared with others. Here, to address this limitation, we trained multi-region recurrent neural networks (RNN) models to reproduce the dynamics of large-scale single-unit recordings (more than 6000 neurons across 7 cortical areas) from monkeys engaged in a two-dimensional (color and motion direction) context-dependent decision-making task. Decoding analyses show that all areas encode both stimuli (color and direction). However, using our approach we uncovered feed-forward and feedback interactions within a network of 7 interacting regions. Constraining interactions during training or testing recovered the canonical brain hierarchy that differentiate sensory and frontal regions. Inspecting across-region interactions, we also found that frontal regions compress the irrelevant stimulus in a context-dependent manner, while sensory regions always compress the same stimulus.**

**Keywords:** neural networks; context-dependent decision making; RNN; large-scale recordings

### Context-dependent decision-making task.

We analysed a previous published dataset (Siegel et al., 2015), where two monkeys were trained on a context-dependent decision making task to categorize either the color (red versus green) or the direction (up versus down) of a colored visual motion stimulus, depending on the context that was cued trial-by-trial (**Fig. 1**). For the purpose of this project, we focused on conditioned-averaged activity, leading to a data tensor $X$ with the shape $C \times T \times N$ , with $C$ = 4 colors $\times$ 4 directions $\times$ 4 context cues (64) conditions, $T$ = 40 time bins of 25 ms and $N$ = 6000 neurons. As extra pre-processing steps, we denoised the data using PCA (Mante & Susillo et al., 2013) and z-scored the activity across conditions. Linear decoding (svm) from this dataset showed that all areas encoded both stimuli (**Fig. 1b**), regardless of the context (Siegel et al., 2015). While there were some quantitative differences in the timing of encoding of each variable, both stimuli were very quickly encoded everywhere.

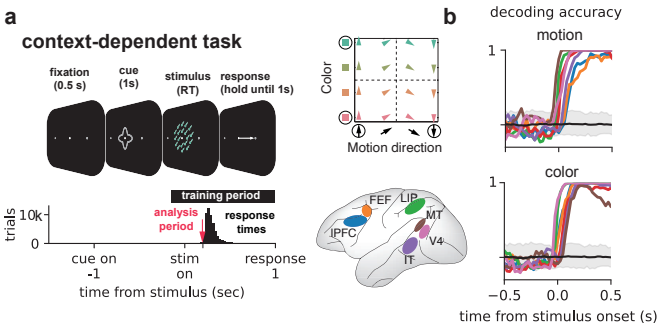


Figure 1: **a)** Experimental design from (Siegel et al., 2015). Red triangle highlights stimulus period used for analyses reported in d-g. Circles in bottom left highlight stimulus used in d-f. **b)** Decoding of color and direction is very high in all areas.

### Multi-region RNN and fitting procedure.

Instead of handcrafting specific across-region interactions (e.g. Barbosa et al., 2021, 2022), we inferred directed interactions between all areas by training rate-based recurrent neural networks with back-propagation to replicate the recorded neural activity of each neuron (Valente et al., 2022). Specifically, we enforced a one-to-one mapping between the recorded and simulated neurons while minimizing the reconstruction error $\mathcal{L}$:

$$\mathcal{L} = \sum_{c=1}^{C} \sum_{i=1}^{N} \sum_{t=1}^{T} [\mathcal{X}_{cit} - x_i^c(t)]^2 \tag{1}$$

$$\dot{x}_i(t) = -x_i(t) + \sum_j J_{ij}\phi(x_j(t)) + \sum_l^{N_{in}} u^l(t) I_i^l + \xi. \tag{2}$$

Here $J$ represents the network connectivity matrix and $\phi$ is $tanh$. In addition to inputs from other neurons, scaled by $J$, each neuron receives feed-forward $N_{in}$ input signals $u^l(t)$ via the weights $I^l$ and independent noise $\xi \sim \mathcal{N}(0, 0.1)$. For the task modelled here, the network received 6 inputs (color, stimulus and 4 different cues) which were delivered to all neurons, except to those in PFC and FEF.

After fitting, we performed two types of experiments, that we detail below. Namely, we blocked across-region interactions (during testing or training) and partitioned across-region inputs.

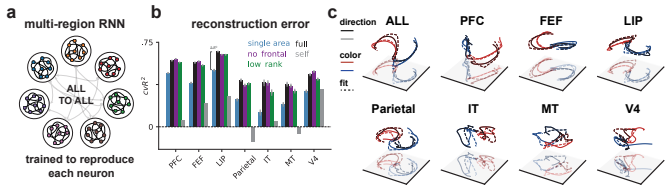


Figure 2: **a)** Illustration of multi-region RNN and **b)** cross-validated reconstruction error for different baselines. **c)** First 3 PCs of original data and model fits, separatedly for each region.

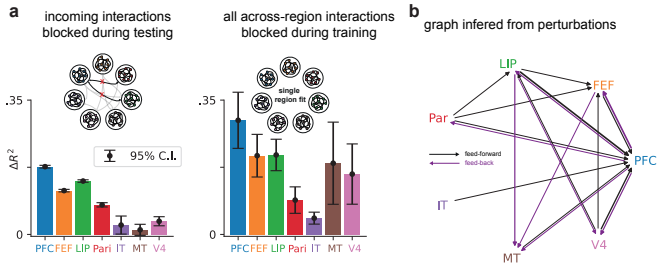


Figure 3: **a)** Difference in reconstruction error ($\Delta R^2$) when blocking across-region communication during testing (left) or training (right). **b)** Visualization of strongest 50% of all the pairwise $\Delta R^2$. Line trickiness represent larger impact.

## Blocking of across-region communication.

We blocked across-region communication during testing and training and compared the full model with the perturbed model (**Fig. 3a**). Removing incoming interactions from all regions using both approaches revealed the canonical brain hierarchy: more frontal regions (e.g. PFC, FEF) integrate information globally, in contrast to sensory regions seem to integrate less (e.g. MT, V4). We also visualized this result during testing in a communication graph **(Fig. 3b)** by blocking communication between each possible pairs of regions and comparing the impact on the model predictability ($\Delta R^2$). In **(Fig. 3b)**, we show the cases where we saw the strongest (mean-split) impact in blocking across-region communication.

## Partitioning inputs.

In this particular analyses, we ran the model forward 10 times, each for 2 colors $\times$ 2 directions $\times$ 2 context (8) conditions and saved each unit's activation in 10 data tensors $X$, with shape $8 \times T \times N$. Note that we trained the model on all the conditions (cross-validated), but for easier visualisation in this analyses we ran the model forward for a subset of 2 stimuli, corresponding to the extreme values (circles in **Fig. 1 a**). In what follows, we report the analyses performed at $t = 0.2s$, a period during stimulus presentation when the animal did respond yet (red arrow in **Fig. 1a**), resulting in a $C \times N$ activity matrix. Using an approach similar to (Perich et al., 2020), we then partitioned the inputs from recurrent interactions (within-area) from those resulting from across-area interactions. Specifically, to estimate the inputs from a source to a target region ($s \to t$), we projected the source area activity $X_{source}$ onto the connectivity block of $J$ corresponding to the interactions between both regions $J_{s \to t}$, resulting in the projection $X_{s \to t} = J_{s \to t}^T X_{source}$. For each source-target pair we considered 4 projections: $X_{source}^{color}$, $X_{source}^{direction}$, $X_{s \to t}^{color}$ and $X_{s \to t}^{direction}$, corresponding to within- and across-area activity projections and for each context separately. To visualize these high-dimensional projections, we plotted their first two principal components (**Fig. 4a**). Visually inspecting these 2D plots, we observed that sensory areas (V4, MT and IT) projected only one variable (color or direction) while compressing others, irrespective of the context or downstream area (see **Fig. 4a**, left for two examples). In contrast, we observed that the prefrontal cortex (PFC) and frontal eye fields (FEF) projected different information depending on the downstream area or context (see **Fig. 4a**, for two examples). This suggested that PFC/FEF compressed the irrelevant stimulus in their projection to frontal areas but not as much towards sensory areas. In the following we describe how we quantify these observations in the original high-dimensional spaces.

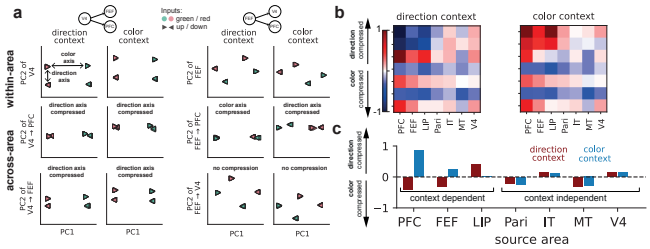


Figure 4: **a)** First two principle components of selected interactions between V4, FEF and PFC illustrated on the top. On the left, it suggests that V4 compresses direction in its interactions. On the right, FEF shows context-dependent (with PFC) and area-dependent compression. **b)** Compression ratio for all source-target pairs and both contexts. **c)** Averaging compression ratios across target areas, suggests that sensory areas always compress their non-preferred stimulus (e.g. MT compresses color, V4 compresses direction), while frontal areas show context-dependent compression of the irrelevant stimulus. Preliminary analyses show that the level of context-dependent compression depends on which areas receive direct stimuli inputs (not shown)

## Compression of stimulus information.

We used Linear discriminant analysis (LDA) to quantify the amount of information related to color and direction separately, and from each of the four projections described above, resulting in 8 decoding values. For example, $D_{s \to t}^{direction}(color)$ for the decoding of color from the cross-area activity subspace (from $s$ to $t$) during the direction context. We then devised a compression metric by comparing the decoding before and after a communication layer. For example, for the compression of color information during the color context and due to the interaction $s \to t$:

$$C_{s \to t}^{direction}(color) = \frac{D_{source}^{direction}(color)}{D_{s \to t}^{direction}(color)} \quad (3)$$

Using this metric, we found that information was compressed in all projections for all stimuli and during both contexts (not shown). To see if compression was biased for some stimuli or context, we devised a final metric that we called compression

ratio:

$$\frac{[C_{s\to t}^{context}(color) - C_{s\to t}^{context}(direction)]}{[C_{s\to t}^{context}(color) + C_{s\to t}^{context}(direction)]} \quad (4)$$

Assuming compression values are positive (which they empirically turned out to be; not shown), this ratio varies between -1 and 1. More negative values show that direction is more strongly compressed by the $s \to t$ layer in the considered context, and positive values that color is more strongly compressed in that same context. In **Fig. 4b**, we report this ratio for all area pairs. As suggested in the selected examples (**Fig. d**), PFC and FEF show context-dependent projections towards frontal-parietal areas (PFC, FEF, LIP, Parietal) but less so for sensory areas. On the other hand, sensory areas compress their non-preferred, regardless of the context.

## References


Barbosa, J., Babushkin, V., Temudo, A., Sreenivasan, K. K., & Compte, A. (2021). Across-area synchronization supports feature integration in a biophysical network model of working memory. *Frontiers in Neural Circuits*, *15*, 716965.

Barbosa, J., Proville, R., Rodgers, C. C., DeWeese, M. R., Ostojic, S., & Boubenec, Y. (2023). Early selection of task-relevant features through population gating. *Nature Communications*, *14*(1), 6837.

Mante, V., Sussillo, D., Shenoy, K. V., & Newsome, W. T. (2013, November). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, *503*(7474), 78–84.

Perich, M. G., & Rajan, K. (2020, December). Rethinking brain-wide interactions through multi-region 'network of networks' models. *Curr. Opin. Neurobiol.*, *65*, 146–151.

Semedo, J. D., Zandvakili, A., Machens, C. K., Yu, B. M., & Kohn, A. (2019, April). Cortical areas interact through a communication subspace. *Neuron*, *102*(1), 249–259.e4.

Siegel, M., Buschman, T. J., & Miller, E. K. (2015, June). Cortical information flow during flexible sensorimotor decisions. *Science*, *348*(6241), 1352–1355.

Valente, A., Pillow, J. W., & Ostojic, S. (2022). Extracting computational mechanisms from neural data using low-rank rnns. *Advances in Neural Information Processing Systems*, *35*, 24072–24086.