Biologically informed cortical models predict optogenetic perturbations

Christos Sourmpis, Carl Petersen, Wulfram Gerstner, and Guillaume Bellec

Abstract

A recurrent neural network fitted to large electrophysiological datasets should reveal the chain of cortical information transmission. In particular, successful network reconstruction methods should enable a model to predict the response to optogenetic perturbations. We test recurrent neural network (RNN) fitted to electrophysiological datasets on unseen optogenetic interventions, and we propose an RNN extraction method that better predicts the response of the biological circuit to these perturbations. Our method uses gradient descent optimization to reproduce various aspects of neuronal activity patterns, and introduces biologically informed inductive biases like spiking neuron dynamics and structured connectivity reflecting excitatory and inhibitory cell types. We show that including these biological inductive biases can drastically improve the robustness of optogenetic perturbations compared to sigmoidal RNN models used predominantly in the field. Furthermore, we show that gradients of the fitted RNN can predict the effect of micro-perturbations in the recorded circuits, and discuss potentials for measuring brain gradients and designing experimental setups with gradient-targeted stimulation to bias an animal behavior.

1 Introduction

A fundamental question in neuroscience is how cortical circuit mechanisms drive perception and behavior. To tackle this question, the field has been collecting large-scale electrophysiology datasets under reproducible experimental settings (Siegle et al., 2021; Esmaeili et al., 2021; Urai et al., 2022; International Brain Laboratory et al., 2023). However, the field lacks data-grounded modeling approaches to generate and test hypotheses on the causal role of neuronal and circuit-level mechanisms. To leverage the high information density of contemporary recordings, we need both (1) modeling approaches that scale well with data, and (2) metrics to quantify when models improve our understanding of the observed phenomena.

Biophysical simulations have been crucial for our understanding of single-cell mechanisms (Hodgkin, 1958), and have been used to describe interactions across cortical layers, columns, and areas (Markram et al., 2015; Billeh et al., 2020; Isbister et al., 2023). A promising approach to constrain models to electrophysiology recordings approach lies in the optimization of the simulation parameters by gradient descent. These methods were successful in quantitatively classifying functional cell types (Pozzorini et al., 2015; Teeter et al., 2018), and modeling micro-circuit interactions (Pillow et al., 2008; Deny et al., 2017; Mahuas et al., 2020). To bridge the gap from single neurons or small retinal networks to cortical recordings, recent studies made substantial progress towards data-constrained recurrent neural network (RNN) models (Perich et al., 2020; Bellec et al., 2021; Arthur et al., 2023; Kim et al., 2023; Dinc et al., 2023; Sourmpis et al., 2023; Pals et al., 2024). In this line of work, neurons in the RNN are mapped "one-to-one" to recorded cells and optimized by gradient descent to predict recorded activity at large.

An important question is whether these data-constrained RNN can reveal a truthful mechanism of neuronal activity and behavior. By construction, the RNNs can generate

brain-like network activity, but how can we measure whether the reconstructed network faithfully represents the biophysical mechanism? To answer this question, we submit a range of RNN reconstruction methods to a difficult *perturbation test*: we measure the similarity between the response to unseen perturbations in the RNN and the recorded biological circuit.

Optogenetics is a powerful tool to induce precise causal perturbations in vivo (Esmaeili et al., 2021; Guo et al., 2014). It involves the expression of light-sensitive ion channels (Aravanis et al., 2007), such as channelrhodopsins, in specific populations of neurons (e.g., excitatory/pyramidal or inhibitory/parvalbumin-expressing). In this paper we use datasets combining dense electrophysiological recordings with optogenetic perturbations to evaluate RNN reconstruction methods. Since the cells are mapped one-to-one to recorded neurons for all the RNNs that we consider, we can induce optogenetic perturbations targeting cell-types and areas as done in vivo. Yet, we observe that the similarity between the simulated and recorded perturbations vary greatly depending on the reconstruction methods.

Most prominently, we study two opposite types of RNN specifications. First, as a control model, we consider a traditional sigmoidal RNN (σRNN) which is arguably the most common choice for contemporary data-constrained RNNs (Perich et al., 2020; Arthur et al., 2023; Pals et al., 2024); and second, as an alternative, we develop a model with biologically informed inductive biases (bioRNN): (1) neuronal dynamics follow a simplified spiking neuron model, and (2) neurons associated to fast-spiking inhibitory cells have short inhibitory projections while other neurons are excitatory. Following Neftci et al. (2019); Bellec et al. (2018b); Sourmpis et al. (2023), we adapt gradient descent techniques to optimize the bioRNN parameters of neurons and synapses to explain the recorded neural activity and behavior.

Strikingly, we find that the bioRNN is more robust to perturbations than the σRNN. This is nontrivial because it is in direct contradiction with other metrics used often in the field: the σRNN achieves higher similarity with unseen recorded trials before perturbation. This contradiction is confirmed both on synthetic and in vivo datasets. To analyze this result, we submit a spectrum of intermediate bioRNN models to the same *perturbation tests*, and identify two bioRNN model aspects which are most important to improve: (1) Dale's law (the cell type constrains the sign of the connections), and (2) local-only inhibition (inhibitory neurons do not project to other cortical areas). We also find that other biological inductive biases like spiking neuron dynamics and the sparsity prior improve the robustness to perturbation but to a lesser extent.

Furthermore, a perturbation-robust bioRNN enables the prediction of the causal effect of perturbations in the recorded circuit. This becomes particularly interesting with micro-perturbations (μ-perturbation) targeting dozens of neurons in a small time windows. We show in silico that the causal effect of μ-perturbations can be well approximated by the RNN gradients, which has two important implications for experimental neuroscience: (1) we can use RNN gradients to target a subset of cells to optimally increase (or decrease) movement in a simulated mouse (this is demonstrated in silico in a prospective experimental setting with fine-grained read-write electrophysiology); (2) since the methodology is robust to perturbations, the RNN gradients enable measurements of the gradient of the real circuit from passive spike train recordings. So fitting RNNs become a tool to measure "brain gradients" and potentially relate contemporary in vivo measurement to decades of theoretical results from machine learning where the gradient is a foundational concept (LeCun et al., 2015; Richards and Kording, 2023).

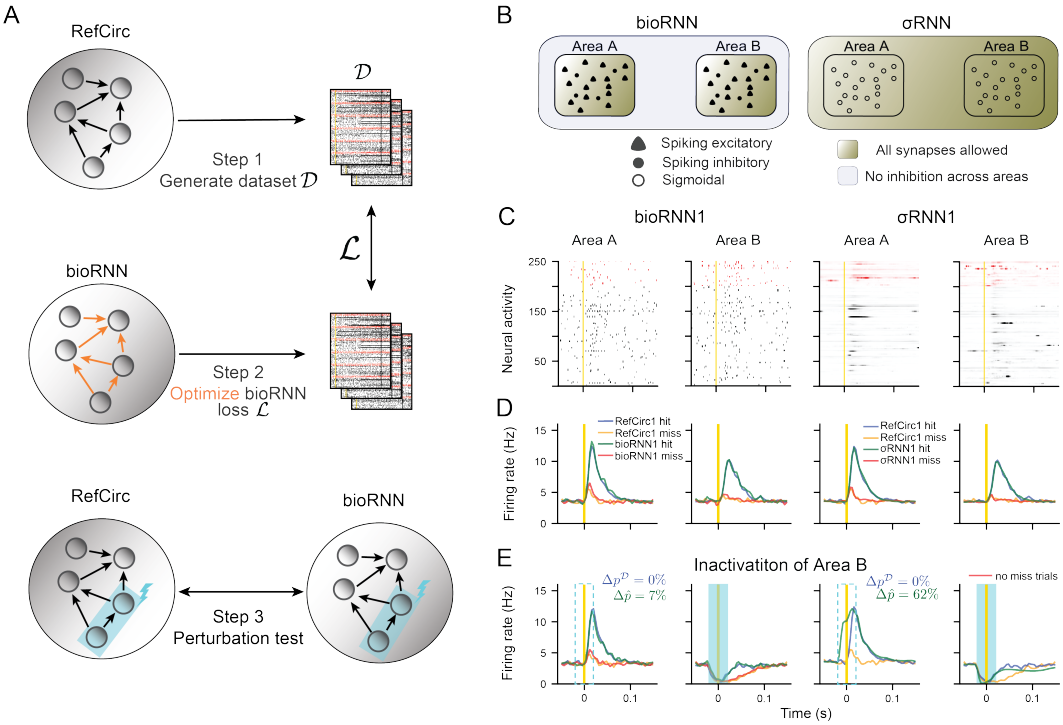

Figure 1: **Network reconstruction and perturbation tests. A.** The three steps to reconstruct the reference circuit (RefCirc) using a biologically informed RNN (bioRNN) and evaluate the reconstruction based on perturbation tests. **B.** Summary of the differences between a bioRNN and a standard sigmoidal network σRNN. **C.** Example raster plot for bioRNN1 and σRNN1, neurons mapped as reference inhibitory cells are shown in red. **D.** Trial-averaged activity of the two circuits during hit (blue: RefCirc1; green: bioRNN1/σRNN1) and miss (yellow: RefCirc1; red: bioRNN1/σRNN1) trials. A trial is a hit if both areas reach a transient firing rate above 8Hz, otherwise it is miss. All models display a hit trial frequency of $p \approx 50\%$. **E.** Same as **D** during inactivation of area B. Neurons directly targeted by the light are excluded from the averaging for the hit trial classification. Δp is the change of hit trial frequency. The groundtruth $\Delta p^{\mathcal{D}}$ is 0% for the feedforward reference circuit RefCirc1.

2 Results

2.1 Reconstructed Networks: biological inductive biases strengthen robustness to perturbations

Synthetic dataset for challenging causal inference We build a simple synthetic dataset as the first benchmark to evaluate quantitatively our network reconstruction methods. Let's consider two areas A and B which are either transiently active together ("hit trial" occurring with frequency p) or quiescent together ("miss trial" occurring with probability $1 - p$). Since the two areas are active or in-active together, it is hard to infer if they are connected in a feedforward or recurrent fashion. We prove in Method 4.1 that it is impossible to discard a feedforward or recurrent hypothesis without a closer look at the individual neuron activity. Our toy example therefore represents a common sub-problem that is often encountered when analyzing large cortical recordings of multiple recordings with a low dimensional instructed behavior (Steinmetz et al., 2019; Esmaeili et al., 2021; International Brain Laboratory et al., 2023).

To generate artificial spike train recordings which capture this problem, we design RefCirc1 and RefCirc2, corresponding to the feedforward and recurrent hypotheses, respectively. The RefCircs consist of two populations of 250 spiking neurons each representing putative areas A and B. In each population, 80% of the neurons are excitatory, and the others are inhibitory. In response to a stimulus, the two circuits display the same bi-modal hit-or-miss response. With a hit trial frequency $p = 50\%$: areas A and B are near-simultaneously active (hit trials) or inactive (miss trials). In this synthetic dataset, a trial is considered a hit when the maximum firing rate of the population average in both areas is above 8 Hz. To make the distinction between the feedforward and the recurrent RefCircs even more difficult, each neuron in RefCirc1 has the identical trial-averaged activity as the corresponding neuron in RefCirc2. Fig. S1 shows recorded activity from the two reference circuits and Methods 4.4 explains how to construct the artificial circuits. Essentially, the only difference between the two RefCircs is not visible in the peristimulus time histogram (PSTH) but only in the structure of the network: the feedforward circuit RefCirc1 has strictly zero feedback connections from B to A, whereas the recurrent circuit RefCirc2 has feedback connections from B to A.

The difference between RefCirc1 and RefCirc2 becomes visible with optogenetic perturbations. To simulate optogenetic inactivations of an area, we inject a transient current into the inhibitory neurons, modeling the opening of light-sensitive ion channels (symmetrically, an optogenetic activation is simulated as a positive current injected into excitatory cells). Figure S1B-D shows that activity in response to optogenetic perturbations in area B reflects the presence or absence of feedback connections. Our ultimate *perturbation test* will be to compare the effect of optogenetic perturbations in the reconstructed RNN and in the reference circuits.

Network reconstruction methodology (synthetic dataset) To reconstruct the recorded circuits with an RNN, we record activity from the spiking RefCircs, and optimize the parameters of an RNN to generate highly similar network activity. The whole reconstruction method is summarized in panel A of Figure 1. In the simplest cases, the RNN is specified as a sigmoidal network model (Rosenblatt, 1960; Elman, 1990): σRNN1 and σRNN2 are optimized to reproduce the recording from RefCirc1 and RefCirc2 respectively. In this synthetic dataset, the reconstructed σRNNs have the same size as the RefCirc (500 neurons) and sigmoidal neurons are mapped one-to-one with RefCirc neurons (20% are mapped to inhibitory RefCirc neurons). They are initialized with all-to-all connectivity and are therefore blind to the structural difference of the RefCirc1 and 2 (feedforward or recurrent). From each of the two RefCircs, we store 2,000 trials of simultaneous spike train recordings of all 500 neurons from areas A and B (step 1 in Fig. 1A). Half of the trials form the training dataset and will be the basis for our data-driven RNN optimization. The

second half of the recorded trials form the testing set and are used to evaluate the quality of the reconstruction before perturbations.

We optimize the synaptic "weights" to minimize the difference between the activity of the RefCirc and that of the σRNN (step 2 in Fig 1A, see Methods). The optimization of these parameters minimizes three loss functions: (i) the neuron-specific loss function $\mathcal{L}_{\text{neuron}}$ is the mean-square error of the *trial-averaged* spike counts (e.g. the PSTH) between σRNN and RefCirc neurons. (ii) To account for fluctuations of the single-trial network activity, we use a trial-specific loss function $\mathcal{L}_{\text{trial}}$, which is the distance between the distribution of single trial population-averaged activity of σRNN and RefCirc (we follow the methodology from Sourmpis et al. (2023), where this is called the trial matching loss function). (iii) Finally, we add a regularization loss function $\mathcal{L}_{\text{reg}}$ to penalize large weights. The three losses are defined mathematically in Methods 4.5.

We also developed a biologically informed RNN model (bioRNN) for which we have designed a successful optimization technique. The main differences between σRNNs and bioRNNs are the following. Firstly, the bioRNN neuron model follows a simplified leaky integrate and fire dynamics (see Methods 4.2) yielding strictly binary spiking activity. Secondly, we constrain the recurrent weight matrix to describe cell-type specific connectivity constraints: following Dale's law, neurons have either non-negative or non-positive outgoing connections; moreover, since cortical inhibitory neurons rarely project across areas, we assume our model inhibitory neurons to project only locally, i.e., within the same area. Thirdly, we add a term to the regularization loss $\mathcal{L}_{\text{reg}}$ to implement the prior knowledge that cross-area connections are more sparse than within an area. Adding these biological features into the model requires an adapted gradient descent algorithm and matrix initialization strategies (Methods 4.5). The reconstruction method with σRNNs and bioRNNs is otherwise identical: the models have the same size, and are optimized on the same data, for the same number of steps and using the same loss functions. The two models bioRNN1 and bioRNN2 are optimized to explain recordings from RefCirc1 and Refcirc2, respectively. Importantly, the structural difference between RefCirc1 (feedforward) and RefCirc2 (feedback) is assumed to be unknown during parameter optimization: at initialization, excitatory neurons in bioRNN1 or bioRNN2 project to any neuron in the network with transmission efficacies (aka as synaptic weights) initialized randomly.

After parameter optimization, we have four models, σRNN1, σRNN2, bioRNN1 and bioRNN2, that we will call 'reconstructed' models in the following. To validate the reconstructed models, we verify that the network trajectories closely match the data on the test set in terms of (i) the "behavioral" hit-trial frequency (a trial is a "hit" when the population average firing rate is higher than 8Hz), (ii) the peristimulus time histogram (PSTH) mean-square error of single neurons as evaluated by $\mathcal{L}_{\text{neuron}}$, and (iii) the distance between single-trial network dynamics as evaluated by $\mathcal{L}_{\text{trial}}$ (see Suppl. Fig. S2 and Suppl. Table 1). At first sight, the σRNN display a better performance when comparing the activity simulated by the RNN and recordings held out in the testing set: the distance of single trial network dynamics is lower with σRNN (this metric is equal to $\mathcal{L}_{\text{trial}}$, see all measurements in Suppl. Table 1). This is expected considering that the optimization of bioRNNs is less flexible and efficient because of the sign-constrained weight matrix and the imperfect surrogate gradient approximation through spiking activity. However, the two bioRNNs are drastically more robust when evaluating the models with *perturbation tests*.

Perturbation test To test which of the reconstructed RNNs capture the causal mechanisms of the RefCircs, we simulate optogenetic activations and inactivations of area B (step 3 in Fig. 1A). We first compare the change of hit probability after perturbations in the reconstructed RNN ($\Delta\hat{p}$) and in RefCirc ($\Delta p^{\mathcal{D}}$) in Figure 2. For the σRNN the activation or inactivation of area B changes drastically the peak firing rate in area A: all trials become a hit during inactivation of area B. This is not consistent with the RefCircs where the effect of the perturbation is milder: the firing rate distribution remains bi-modal (hit versus miss) after perturbation with a moderate change of hit probability. By design, there is even

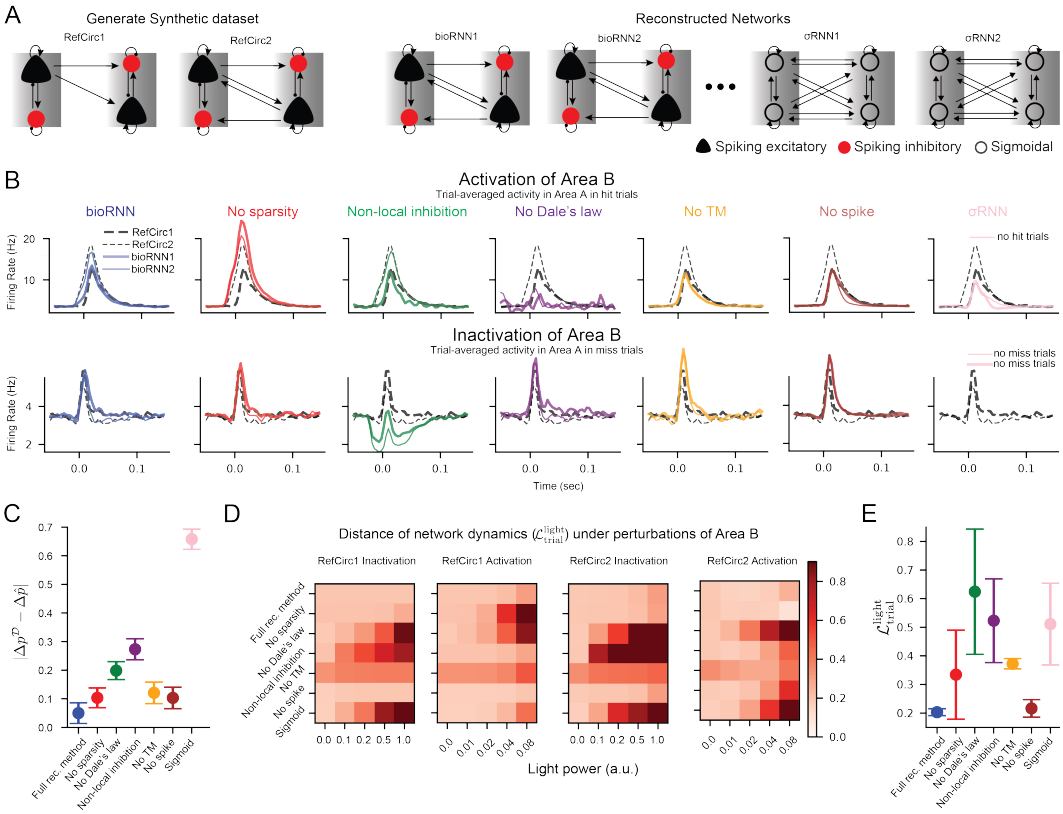

Figure 2: **Biologically informed RNNs predict optogenetic perturbations. A.** Parameters of bioRNN1, bioRNN2, σRNN1 and σRNN2 are optimized to reconstruct RefCirc1 or RefCirc2 from spiking recordings. The bioRNNs and σRNNs are blind to the structural difference of RefCirc1 and 2 and must infer this from the spiking data alone. BioRNN variants are defined by removing one of the biologically inspired features, for instance "No Dale's law" refers to a bioRNN without sign constraints in the weight matrix. **B.** Trial-averaged activity in area A under activation/inactivation of area B. Dashed black lines indicate the activity of RefCirc1 (thick dashed) and RefCirc2 (thin dashed). All the RNNs are tested with the same reference circuit and training data, each bioRNN model variant is shown with a different color. **C.** Error between the change of hit probability after perturbations in the RNN Δp and in the RefCirc Δp. **D.** The distance of network dynamics $\mathcal{L}_{\text{trial}}^{\text{light}}$ between each RNN and RefCirc, as a function of perturbation strength applied to area B (horizontal axis: light power in arbitrary units). **E.** Same quantity as D but averaged for each RNN under all perturbation conditions of area B (activation/inactivation across all power levels).

strictly no change of hit-frequency $\Delta p^{\mathcal{D}} = 0\%$ for RefCirc1 because of the absence of feedback connections from B to A. The bioRNN models capture these changes more accurately (see Fig. 1 and 2). Quantitative results are summarized in Fig. 2C, the error of hit probability changes $|\Delta\hat{p} - \Delta\hat{p}^{\mathcal{D}}|$ is 5% with bioRNNs when averaged over all conditions (bioRNN1 and bioRNN2, with optogenetic inactivations and activations). The corresponding error is 65.8% on average for σRNNs.

To analyze which features of bioRNN explain this robustness to perturbation, we then derive a family of models where only one feature of the reconstruction is omitted. Namely, the "No Dale's law" model does not have excitatory and inhibitory weight constraints, the "Non-local inhibition" model allows inhibitory neurons to project outside of their areas, the "No Spike" model replaces the spiking dynamics with a sigmoidal neuron model, and the "No Sparsity" model omits the cross-area sparsity penalty in $\mathcal{L}_{\text{reg}}$. Omitting all these features in bioRNN would be equivalent to using a σRNN. The accuracy metrics on the testing sets before perturbation are reported for all RNN variants on the Suppl. Fig. S2C and F. For reference, we also include the model "No TM" (trial-matching) which omits the loss function $\mathcal{L}_{\text{trial}}$ during training.

Since the optogenetic perturbation targets cells of a specific cell type, we conjectured that modeling cell-type connection properties (Dale's law and local inhibition) were primarily relevant for robust network reconstruction. We first analyze the "No Dale's law" models having strictly no sign constraints. Lifting the sign constrain is detrimental to the robustness to optogenetic perturbation, as seen in Figure 2: empirically, the hit-rate under perturbation of area B in this variant is largely overestimated (Figure 2C) and the perturbed PSTH in area A is distant from the reference (purple curve in Figure 2B). A different failure arises in the "Non-local inhibition" variant which has different matrix constraints: inhibitory neurons are signed as in the bioRNN but project to all areas. The "Non-local inhibition" model exhibits a reduction in the baseline firing rate of area A (green trace during inactivation in Figure 2B). This reduction can be explained by the mono-synaptic inhibitory connections from area B onto area A that exist in this "Non-local inhibition" model and not in the references RefCirc and the bioRNNs. The strong difference in robustness between bioRNN and the "No Dale's law" and "Non-local inhibition" models illustrate the biological inductive biases of bioRNN that are most relevant to predict the specific optogenetic perturbation experiments. This simulation also proves that the optimization is not capable of compensating for the absence of connectivity constraints: one could imagine that hundreds of recorded trials are enough to infer from our in silico recordings which cell is excitatory or inhibitory, or whether a neuron projects to other areas, but it is not the case here.

It is less clear why spiking activity is immediately relevant to model optogenetic perturbations, however we do find that the spiking bioRNN is also slightly more robust to perturbation than the "No spike" bioRNN variant (see both hit-frequency in Figure 2C and distances of network dynamics $\mathcal{L}_{\text{trial}}$ 2E). This is nontrivial because surrogate gradients Neftci et al. (2019) through binarized spiking activity are less optimization-friendly than exact gradient descent in sigmoidal neurons: the σRNN models achieved consistently better performance before perturbations (see the testing set loss value $\mathcal{L}_{\text{trial}}$ in Suppl. Table 1). This trend is inverted after perturbations, even the optimized loss function $\mathcal{L}_{\text{trial}}$ becomes lower on the perturbed trials with the spiking bioRNN (see Suppl. Table 1 and Figure 2E).

So far, we have identified biological inductive biases of bioRNN that improve its robustness to perturbation. We have not yet analyzed if these features are sufficient to recover the structural difference between the feedforward RefCirc1 and recurrent RefCirc2. To reveal this difference in the optimized bioRNN1 and bioRNN2 models, we find the sparsity prior penalizing the connections across areas to be crucial. This is visible in Figure 2B when comparing the bioRNN with the "No Sparsity" model (blue and red curves). While the "No Sparsity" variant is quantitatively robust to the perturbation (Figure 2C and E), it does not display in the PSTH of hit trials (panel B) the early increase in area A characteristic of the perturbation propagating through the feedback from area B in the recurrent RefCirc2. On

this dataset, only the model with sparsity prior, constrained weights and spiking activity was able to capture this characteristic difference between feedforward and recurrent circuits RefCirc1 and RefCirc2. We will now investigate whether the same approach can predict optogenetic perturbations with data recorded in vivo.

2.2 Predicting perturbations on in vivo electrophysiology data

To test whether our reconstruction method with biological inductive biases can predict optogenetic perturbations in large-scale recordings, we used the dataset from Esmaeili et al. (2021). In this study, water-deprived mice were trained to perform a whisker tactile detection task. In 50% of the trials (Go trials), a whisker is deflected, followed by a 1-second delay, after which an auditory cue signals the mice can lick a water spout to receive a water reward. In the other 50% of trials (No-Go trials), no whisker deflection occurs, and licking after the auditory cue results in a penalty with an extended time-out period. While the mice performed the task, experimenters recorded $6,182$ units from 12 areas across 18 mice. Using this dataset, we focused on the 6 most relevant areas. From each area, we randomly selected 250 neurons (200 putative excitatory and 50 putative inhibitory), which correspond to 1500 neurons in total. These areas, all shown to be causally involved in the task (Esmaeili et al., 2021), include the primary and secondary whisker sensory cortex (wS1, wS2), the primary and secondary whisker motor cortex (wM1, wM2), the anterior lateral motor cortex (ALM), and the primary tongue-jaw motor cortex (tjM1). We fit the neuronal activity using the same reconstruction method as used for the synthetic dataset. To add a simulation of the behavioral output along with the network activity, we simulate the jaw movement of the mouse as a linear readout driven by the model's neural activity. This readout is regressed with the real jaw movement extracted from video footage. The parameter optimization of the behavioral readout is performed jointly with fitting the synaptic weights to the neuronal recordings, see Methods 4.5. After training, our reconstructed model can generate neural activity with firing rate distribution, trial-averaged activity, single-trial network dynamics and behavioral outcome which are all consistent with the recordings (see Suppl. Fig S4). Before perturbations, we observe again that the σRNN model fits the testing set data better than the bioRNN model (see Table 2 and Fig. S4).

We then submit the reconstructed σRNNs and bioRNNs models to *perturbation tests*. These perturbations are acute spatiotemporal optogenetic inactivations of each area during different time periods (see Figure 3B). As an example, we show the effect of an inactivation of wS1 during the whisker period in the model in Fig. 3. By showing a simulated trial with and without perturbation simulated with the with the same random seed, we show that an early perturbation in wS1 can change a lick decision from hit to miss in the model (Fig. 3C). In the in vivo data, we denote the change of lick trial frequency during Go trials $\Delta p^{\mathcal{D}}$ due to the optogenetic perturbations. The empirical measurement were performed in different periods for each area in Esmaeili et al. (2021) (stimulation, delay, or choice periods). For all areas and time windows we measure the corresponding $\Delta\hat{p}$ in the model. On average, the error change probability obtained with the σRNN model is $|\Delta p^{\mathcal{D}} - \Delta\hat{p}| = 22\%$ which is substantially worse than the bioRNN model's 17% (see Figure 3D). As in the synthetic dataset, we find this to be consistent over multiple bioRNN model variant, and we find that imposing Dale's law and local inhibition best explain the improvement in perturbation-robustness. We also measure that the spiking bioRNN predicts the change in lick probability slightly better than the "No Spike" bioRNN model. Conversely, adding the sparsity prior does not seem to improve the perturbed hit-rate prediction on the real data which suggests that the sparsity of connections is not necessary to explain the perturbations of this in vivo dataset.

To further analyze the consistency of the perturbations in the model, we can compare the perturbation map showing changes in lick probability obtained from acute inactivation in the data and the model. The Suppl. Fig. S5 summarizes visually which area has a critical role at specific time points. The change of lick probability in area wS1, ALM and tjM1 are

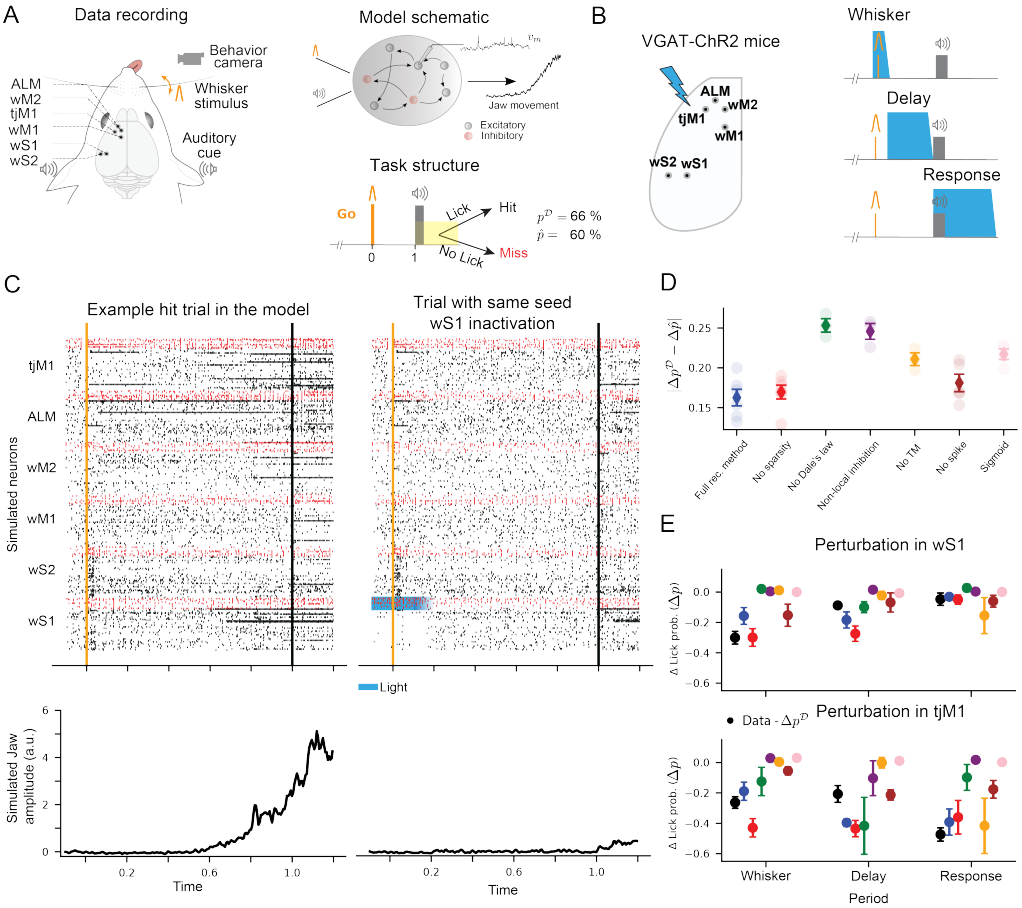

Figure 3: **Predicting optogenetic perturbations for in vivo electrophysiology data** **A.** During a delayed whisker detection task, the mouse reports a whisker stimulation by licking to obtain a water reward. Jaw movements are recorded by a camera. Our model simulates the jaw movements and the neural activity from six areas. **B.** The experimentalists performed optogenetic inactivations of cortical areas (one area at a time) in three temporal windows. **C.** Example hit trial of a reconstructed network (left). Using the same random seed, the trial turns into a miss trial if we inactivate area wS1 (right, light stimulus indicated by blue shading) during the whisker period by stimulation of inhibitory neurons (red dots). **D.** Error of the change in lick probability caused by the perturbation, $\Delta\hat{p}$ is predicted by the model, and $\Delta p^{\mathcal{D}}$ is recorded in mice. Light-shaded circles show individual reconstructed networks with different initializations. The whiskers are the standard error of means. **E.** Examples of $\Delta\hat{p}$ probability changes under perturbation for wS1 (Top) and tjM1 (Bottom). See Suppl. Fig. S5 for the other areas.

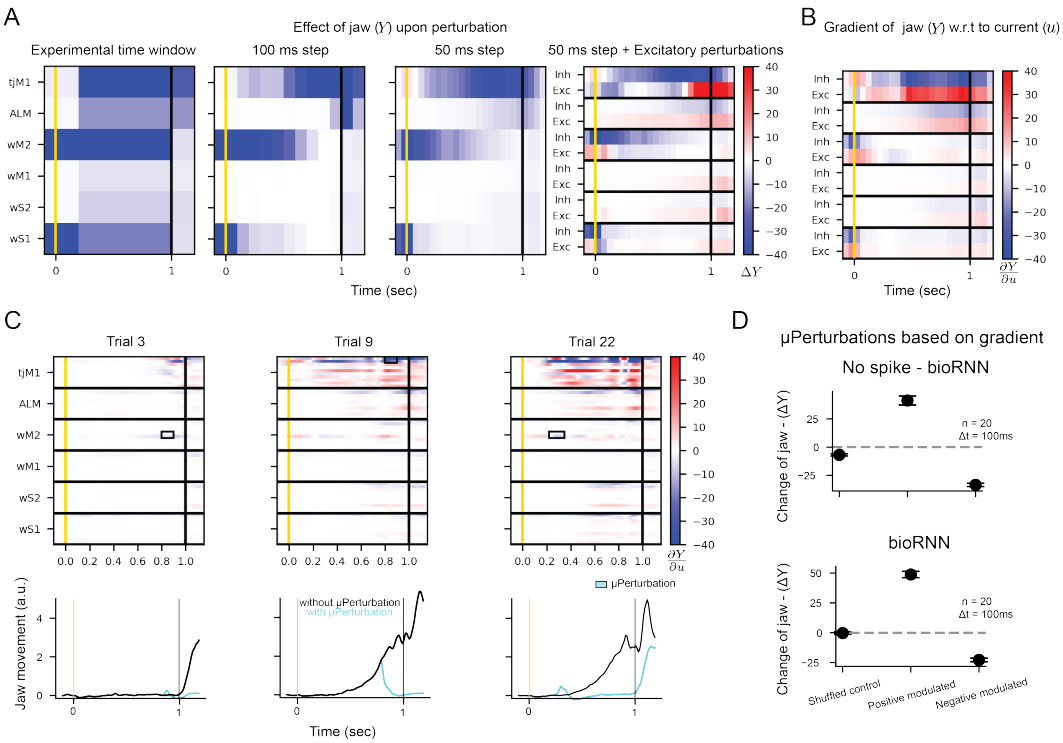

Figure 4: **Measuring circuit gradients with μ-perturbations A-B.** Numerical verification for equation (1). **A** shows the change of jaw movement ΔY following inactivations in a "No Spike" bioRNN. From left to right, we reduce the size of the spatiotemporal window for the optogenetic stimulation. **B.** Gradients values $\sum_{i,t} \frac{dY}{du}$ that approximate ΔY from **A** using equation (1). **C-D.** Verification that gradients predict the change of movement on single trials. In **C**, we stimulate the 20 neurons that have the highest (or lowest) gradient in a 100ms time window, the results are shown in **D**.

accurately predicted by the bioRNN. In contrast, our model tends to underestimate the causal effect induced by the inactivations of wS2, wM1 and wM2 (Suppl Fig S5). Overall, our model is consistent with a causal chain of interaction from wS1 to ALM and continuing to tjM1.

2.3 Applications for experimental electrophysiology

With future progress in recording technology and reconstruction methods, network reconstruction may soon predict the effect of optogenetic perturbation with even higher accuracy. In this section, we explore possible consequences and applications for experimental electrophysiology. In summary, we demonstrate in the following that (1) perturbation-robust bioRNNs enable us to estimate gradients of the recorded circuits, (2) which in turn enable us to target μ-perturbation in the recorded circuit and optimally increase (or decrease) induced movements in our simulated mouse. In the following, the so-called "recorded circuit" is a bioRNN trained on the in vivo dataset that we use as a proxy experimental preparation. Its mathematical unpinning enables for rigorous theoretical considerations and the design of forward-looking in silico experiments.

μ-perturbations measure brain gradients We first prove a mathematical relationship between gradients in the recorded circuit and μ-perturbations. We define the integrated movement as $Y = \sum_t y_t$ where y_t is the movement of the jaw movement at time t generated

by the model, and we denote $\Delta Y^{\lightning}$ as the change of movement caused by the μ-perturbation. If the circuit has well-defined gradients (e.g. say a "No spike" bioRNN model trained on the in vivo recordings in the previous section), using of a Taylor expansion finds that:

$$\Delta Y^{\lightning} = \sum_{i,t \in \mathcal{I}} \frac{dY}{du_i^t} \Delta u_i^t \quad + \epsilon, \tag{1}$$

where $\mathcal{I}$ are the neuron and time indices selected for the optogenetic intervention. The error term ϵ is negligible when the current Δu_i^t induced by the light is small. We confirm this approximation numerically in Fig. 4A: we display movement perturbations $\langle \Delta Y^{\lightning} \rangle$ in the circuit with time windows of decreasing sizes ($\langle \cdot \rangle$ indicates a trial average). When the time window is small, and the perturbation is only applied to excitatory or inhibitory cells in Fig. 4A, one can appreciate visually the similarity with the binned gradient $\langle \sum_{i,t} \frac{dY}{du_i^t} \rangle$ in Fig. 4B. Proceeding to a quantitative verification of equation (1), we now compare the effect of small perturbations targeting only 20 neurons on a single-trial. We use the gradient $\sum_{i,t} \frac{dY}{du_i^t}$ (see Fig. 4C) to predict the outcome of μ-perturbation as follows: for each trial, and each 100ms time window, we identify 20 neurons in the model with highest (or lowest) gradients $\sum_{i,t} \frac{dY}{du_i^t}$. We then re-simulate the exact same trial with identical random seed, but induce a μ-perturbation on selected neurons (see rectangles in Figure 4). If we target neurons with strongly positive gradients, the perturbed jaw movements are strongly amplified $\Delta Y^{\lightning} > 0$; conversely, if we target neurons with negative gradients the jaw movements are suppressed $\Delta Y^{\lightning} < 0$. Although the equation (1) is only rigorously valid for models with well-defined gradients like the "No Spike" model, we also confirm in Fig. 4D that this numerical verification also holds in a spiking circuit model where the gradients are replaced with surrogate gradients (Neftci et al., 2019).

An implication of equation (1) is that the measurements $\langle \Delta Y^{\lightning} \rangle$ that can be recorded in vivo are estimates of the gradients $\langle \sum_{i,t} \frac{dY}{du_i^t} \rangle$ in the recorded circuit. Yet, measuring detailed gradient maps (or perturbation maps) as displayed in Fig. 4 would be costly in vivo as it requires to average $\Delta Y^{\lightning}$ over dozens of trials for each spatio-temporal window. Instead, gradient calculation in a bioRNN model (that was fitted to the experimental preparation) is a rapid mathematical exercise. If the extracted model is valid, then (1) the gradients $\sum_{i,t} \frac{dY}{du_i^t}$ in the bioRNN approximate the effect of μ-perturbations $\Delta Y^{\lightning}$ in the experimental preparation; (2) the gradient $\sum_{i,t} \frac{dY}{du_i^t}$ in the bioRNN approximate the gradient $\sum_{i,t} \frac{dY}{du_i^t}$ in the recorded circuit.

Targeting in vivo μ-perturbations with bioRNN gradients Building on this theoretical finding, we build a speculative experimental setup where the bioRNN gradients are used to target a μ-perturbation and increase (or decrease) the movements Y in the experimental preparation in real time. We show a schematic of this speculative set-up in Fig. 5C for a real in vivo setup and demonstrate in silico in Fig. 5A how to use bioRNN gradients to bias the simulated mouse movement Y. As a preparation step, and before applying perturbations, we assume that the bioRNN is well fitted to the recorded circuit and we collect a large databank $\mathcal{B}$ of simulated trials from the fitted bioRNN. Then in real-time, we record the activity from the experimental preparation until the time t^* at which the stimulation will be delivered (Step 1 in Fig. 5A, t^* is 100ms before the decision period). Rapidly, we find the trial with the closest spike trains in the databank of simulated trials (Step 2) and use the corresponding gradient maps to target neurons with the highest gradient $\frac{dY}{du}$ in the model (Step 3). The targeted stimulation is then delivered immediately at t^* to the experimental preparation (Step 4). When testing this in silico on our artificial experimental preparation, we show in Fig. 5C that this approach can bias the quantity of jaw movement Y driven by the circuit in a predictable way. The amount of movement is consistently augmented if we target neurons with the highest $\frac{dY}{du}$ (or diminished if we target neurons with the lowest $\frac{dY}{du}$).

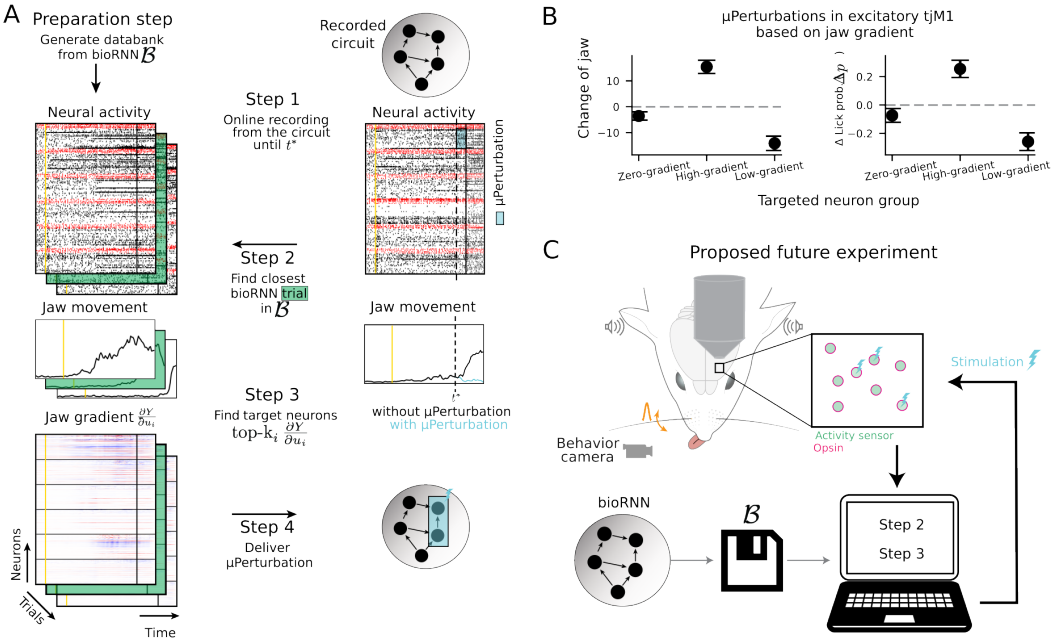

Figure 5: **Gradient targeted μ-perturbations could bias an animal behavior A.** Protocol to deliver an optimal μ-perturbation on the experimental preparation based on jaw gradients. (Step 1) the circuit is recorded until stimulation time t^*. (Step 2) The closest bioRNN trial to the ongoing recorded trial is retrieved from the databank $\mathcal{B}$. (Step 3) We select the neurons with the highest (or lowest) gradient value for the μ-perturbation. (Step 4) The μ-perturbation is delivered at t^*. **B.** Effect of the μ-perturbation using the artificial setup **A** under different light protocols. Practically, for "High gradient", we keep step 3 as it is, for "Low gradient", we change the sign of the gradient, and for "Zero gradient", we pick the 40 neurons with lowest gradient norm. **C.** Speculative schematic of a close-up setup implementing the protocol **A** using read-write optical recordings.

3 Discussion

Finding the right level of detail to model recorded phenomena has sparked intensive debates in computational neuroscience. Our results show that the co-development of *Perturbation tests* and biological modeling provides a quantitative approach to this problem. *Perturbation tests* can detect two types of erroneous brain models: (1) machine learning models with high predictive power, but capturing a wrong causal mechanism; (2) physiological model details that are wrongly claimed to explain measurements in the recorded circuit.

Our reconstruction method uses data-constrained RNN with informed inductive biases, and it is validated on *perturbation tests*. We achieve a reconstruction of the sensory-motor pathway in the mouse cortex during a sensory detection task from electrophysiology data. The model is optimized to explain passive recordings and generalizes better than standard models to in vivo optogenetic interventions. We found unambiguously that anatomically informed sign and connectivity constraints for dominant excitatory and inhibitiory cell types improve the model robustness to optogenetic perturbations. In hindsight, we infer that adding biological constraints becomes beneficial when it is a good trade-off between (1) modeling circuit properties that are relevant for that perturbation (here, the targeted cell types and area-level connectivity properties), (2) not impairing the efficiency of the optimization process. There is a dense knowledge of unexploited details at the connectivity, laminar or cell-type level that could be added into a cortical model like ours (Harris et al., 2019; Liu et al., 2022; Udvary et al., 2022; Staiger and Petersen, 2021). With the relevant *perturbation tests*, extending our approach offers a new quantitative method to measure what are the functional roles of these features. We do not rule out, that significant improvements on *perturbation tests* can also be achieved without any explicit connection with brain physiology (e.g using low-rank connectivity Dubreuil et al. (2022); Valente et al. (2022), or efficient model architectures that have no relation-ship with the brain physiology Azabou et al. (2024); Pandarinath et al. (2018)). However, the failure of a σRNN in our experiments shows the high predictive power of generic models can hide a high sensitivity of perturbations. Significant effort is therefore needed to adapt generic machine learning models towards *perturbation tests* which are out of training distribution by nature. Our conclusion that modeling the biological inductive biases which are most relevant to the perturbations at hand is a reasonable way forward. Broadly speaking this is also supported by other results elsewhere in neuroscience. For instance: biologically inspired topological maps improve convolutional models of the Monkey's visual system Schrimpf et al. (2024), and detailed cell-type distribution and connectome improve vision model in the fly brain Lappalainen et al. (2023); Cowley et al. (2024).

To highlight the importance of perturbation-robust circuit models, we have highlighted implications for experimental neuroscience in section 2.3. We have shown that network reconstructions generate detailed perturbation sensitivity maps from passive recordings only. These sensitivity maps correspond to the RNN gradients and can be used to make targeted micro-stimulation in vivo. More conceptually, we have shown that the gradients of a perturbation robust RNN are also consistent with the gradients of the true recorded circuits. In perspective with the foundational role of gradients in machine learning theory LeCun et al. (2015); Richards and Kording (2023), it enables the measurement of "brain gradients" and lays a computational link between in vivo experimental research and decades of theoretical results on artificial learning and cognition.

4 Methods

4.1 Mathematical toy model of the difficult causal inference between H1 and H2

Let's consider two simplistic mathematical models that both depend on two binary random variables A and B which represent that putative area A is active as $A = 1$ and area B as $B = 1$. With this notation, we can construct two hypothetical causal mechanisms $H1$ ("a feedforward hypothesis") and $H2$ ("a recurrent hypothesis"), which are radically different. The empirical frequency $p(A, B)$ of the outcome does not allow us to differentiate whether the system was generated by a feedforward mechanism $H1$ or a recurrent mechanism $H2$. Schematically, we can represent the two mechanism hypotheses as follows:

$$(H1) \quad A \longrightarrow B, \tag{2}$$

$$(H2) \quad A \longleftrightarrow B\ . \tag{3}$$

For hypothesis $H1$: we assume that external inputs are driving the activity of area A such that $A = 1$ is active with probability p_0, and there are strong feed-forward connections from A to B causing systemically $B = 1$ as soon as $A = 1$. Alternatively, in $H2$, we assume that areas A and B receive independent external inputs with probability $\sqrt{p_0}$. Each of these two inputs is sufficient to cause $A = 1$ or $B = 1$, and the two areas are also strongly connected, so $A = 1$ always causes $B = 1$ and vice versa. Under these hypothetical mechanisms $H1$ and $H2$, the empirical probability table $p(A, B)$ is identical: $p(A = 1, B = 1) = p_0$ ("Hit trial", both areas are active), $p(A = 0, B = 0) = 1 - p_0$ ("Miss trial", the areas are quiescent). In both cases, the possibility that only one area is active is excluded by construction. In other words, even if we observe an infinite number of trials, discriminating the two possible causal interactions between A and B is impossible if we only record whether an area is active or not.

A solution to discriminate between hypotheses $H1$ and $H2$ is to induce a causal perturbation. We can discriminate between our two hypotheses if we can impose a perturbation that forces the inactivation of Area B in both mathematical models. Under the feedforward mechanism $H1$ and inactivation of B, A is not affected $p(A = 1) = p_0$. Under the recurrent hypothesis, $H2$, and inactivation of B, A is activated only by its external input such that $p(A = 1) = \sqrt{p_0}$. This illustrates mathematically how a causal perturbation can be used to discriminate between those two hypothetical mechanisms.

4.2 Neuron and jaw movement model

We model neurons as leaky-integrate and fire (LIF) neurons. The output of every neuron j at time t is a binary outcome z_j^t (spike if $z_j^t = 1$, no spike if $z_j^t = 0$) generated from a membrane voltage v_j^t. The following equations give the dynamics of the membrane voltage v_j^t:

$$v_j^t = \alpha_j v_j^{t-1} + (1 - \alpha_j) u_j^t - v_{\mathrm{thr},j} z_j^{t-1} + \xi_j^t \tag{4}$$

$$u_j^t = \sum_{d,i} W_{ij}^{rec,d} \frac{z_i^{t-d}}{\delta t} + \sum_i W_{ij}^{\mathrm{in}} \frac{x_i^t}{\delta t} \tag{5}$$

where W_{ij}^d, and $W_{ij}^{in,d}$ are the recurrent and input weight matrices. The timestep of the simulation δt is 2 ms when we simulate the real dataset and 1 ms otherwise. The superscript d denotes the synaptic delay; every synapse has one synaptic delay of either 2 or 3 ms. With $\alpha_j = \exp\left(-\frac{\delta t}{\tau_{m,j}}\right)$, we define the integration speed of the membrane voltage, where $\tau_m = 30$ ms for excitatory and $\tau_m = 10$ ms for inhibitory neurons. The noise source

ξ_j^t is a Gaussian random variable with zero mean and standard deviation $\beta_j v_{\text{thr},j}\sqrt{\delta t}$ (β_j is typically initialized at 0.14). The input x_i^t is a binary pulse signal with a duration of 10 ms. For the real dataset, we have two binary pulse input signals, one for the whisker deflection and one for the auditory cue. The spikes are sampled with a Bernoulli distribution $z_j^t \sim \mathcal{B}(\exp(\frac{v_j^t - v_{\text{thr},j}}{v_0}))$, where v_0 is the temperature of the exponential function and $v_{thr,j}$ is the effective membrane threshold. After each spike, the neuron receives a reset current with an amplitude of $v_{trh,j}$ and enters an absolute refractory period of 4 ms, during which it cannot fire.

For networks fitted to the real dataset, we also simulate the jaw movement. The jaw movement trace y is controlled by a linear readout from the spiking activity of all excitatory neurons. Specifically, y is computed as $y = \exp(\tilde{y}) + b$, where b is a scaling parameter and $\tilde{y}^t$ is given by $\tilde{y}^t = \alpha_{jaw}\tilde{y}^{t-1} + (1-\alpha_{jaw})\sum_{d,j} W_j^{jaw} z_j^{t-d}$. Here, W_j^{jaw} is the output weight matrix (linear readout) for the jaw, and $\tau_{jaw} = 5$ms defines $\alpha_{jaw} = \exp(-\frac{\delta t}{\tau_{jaw}})$, which controls the integration velocity of the jaw trace.

4.3 Session-stitching and network structure

As in (Sourmpis et al., 2023), we simulate multi-area cortical neuronal activity fitted to electrophysiology neural recordings. Before we start the optimization, we define and fix each neuron's area and cell type in the model by uniquely assigning them to a neuron from the recordings. For the real dataset from Esmaeili et al. (2021), the cell type is inferred from the cell's action potential waveform (with fast-spiking neurons classified as inhibitory and regular-spiking neurons as excitatory). Most electrophysiology datasets include recordings from multiple sessions, and our method would typically require simultaneous recordings of all neurons. To address this challenge, we introduce a technique called "session-stitching" which allows neighboring modeled neurons to be mapped with neurons recorded across multiple sessions. This effectively creates a "collage" that integrates data from multiple sessions within our model. This approach has practical implications for our optimization process. Specifically, the trial-matching loss includes a term for each session, with the overall loss calculated as the average across all sessions (see 4.5).

For both the real and the synthetic datasets, we simulate each area with 250 LIF neurons and impose that each area has 200 excitatory neurons and 50 inhibitory. Respecting the observation that inhibitory neurons mostly project in the area that they belong (Tamamaki and Tomioka, 2010; Markram et al., 2004), we don't allow for across-area inhibitory connections. The "thalamic" input is available to every neuron of the circuit, and the "motor" output for the real dataset, i.e., jaw movement, is extracted with a trained linear readout from all the excitatory neurons of the network, see 4.2.

4.4 Reference circuits for hypotheses 1 and 2

To build a synthetic dataset that illustrates the difficulty of separating the feedforward (H1) and recurrent hypotheses (H2), we construct two reference spiking circuit models RefCirc1 and RefCirc2. The two networks consists of two areas A and B, and their activity follows the hard causal inference problem from method 4.1, making it hard to distinguish A1 and A2 when recording the co-activation of A and B. Moreover, to make the problem even harder, the two networks are constructed to make it almost impossible to distinguish between H1 and H2 with dense recordings: the two circuits are designed to have the same PSTH and single-trial network dynamics despite their structural difference, one is feedforward and the other is recurrent.

To do so, RefCirc1 and 2 are circuit models that start from random network initializations following the specifications described in Methods 4.2 and 4.3. The only difference is that we do not allow feedback connections from A to B in RefCirc1, the construction below is otherwise identical. The synaptic weights of the two circuits are optimized with the losses described in Methods 4.5 to fit the identical target statistics in all areas: the same

PSTH activity for each neuron and the same distribution of single-trial network dynamics. The target statistics are chosen so the activity in RefCirc1 and 2 resemble kinematics and statistics from a primary and a secondary sensory area. The baseline firing rates of the neurons is dictated by the target PSTH distribution and it follows a log-normal distribution, with excitatory neurons having a mean of 2.9 Hz and a standard deviation of 1.25 Hz and inhibitory neurons having a mean of 4.47 Hz and a standard deviation of 1.31 Hz. The distribution of single-trial activity is given by the targeted single-trial dynamics: in RefCirc1 and 2, the areas A and B respond to input 50% of the time with a transient population average response following a double exponential kernel characterized by $\tau_{rise} = 5$ ms and $\tau_{fall} = 20$ ms. Mimicking a short signal propagation between areas, these transients have a 4 or 12 ms delay in areas A and B respectively (relative to the onset time of the stimulus). To impose a "behavioral" hit versus miss distribution that could emerge from a feedforward and recurrent hypothesis (see method 4.1), the targeted population-averaged response of each trial is either a double-exponential transient in both area A and B ("Hit trials"), or remains at a baseline level in both areas ("Miss trials") in the remaining trials. At the end of the training, we verified that RefCric1 and RefCirc2 generate very similar network activity in the absence of perturbation (see Figure S1). The circuits are then frozen and used to generate the synthetic dataset. We generate 2000 trials from these RefCircs, 1000 of which are used for the training set and 1000 for the testing set.

4.5 Optimization and loss function

The optimization method we use to fit our models is back-propagation through time (BPTT). To overcome the non-differentiability of the spiking function, we use surrogate gradients (Neftci et al., 2019). In particular, we use the piece-wise linear surrogate derivative from Bellec et al. (2018b). For the derivative calculations, we use $\frac{v_j^t - v_{\text{thr},j}}{v_0}$ and not $\exp(\frac{v_j^t - v_{\text{thr},j}}{v_0})$. We use sample-and-measure loss functions that rely on summary statistics, as in (Bellec et al., 2021; Sourmpis et al., 2023), to fit the networks to the data. Our loss function has two main terms: one to fit the trial-averaged activity of every neuron ($\mathcal{L}_{\text{neuron}}$), and one to fit the single trial population average activity ($\mathcal{L}_{\text{trial}}$), $\mathcal{L} = \mathcal{L}_{\text{neuron}} + \mathcal{L}_{\text{trial}}$. The two terms of the loss function are reweighted with a parameter-free multi-task method (Défossez et al., 2023) that enables the gradients to have comparable scales.

As in Sourmpis et al. (2023): (1) To calculate the **trial-averaged loss**, we first filter the trial-averaged spiking activity $\mathcal{T}^t_{\text{neuron},j}(\boldsymbol{z}) = \frac{1}{K}\sum_k \boldsymbol{z}^t_{j,k} * f$ using a rolling average window (f) of 8 ms. We then normalize it by the trial-averaged filtered data activity, $\mathcal{T}'^t_{\text{neuron},j}(\boldsymbol{z}) = (\mathcal{T}^t_{\text{neuron},j}(\boldsymbol{z}) - \langle \mathcal{T}^t_{\text{neuron},j}(\boldsymbol{z}^{\mathcal{D}})\rangle_t)/(\sigma_t(\mathcal{T}^t_{\text{neuron},j}(\boldsymbol{z}^{\mathcal{D}}))$, where $\langle . \rangle_t$ is the time average, and σ_t the standard deviation over time. The trial-averaged loss function is defined as:

$$\mathcal{L}_{\text{neuron}} = \sum_j^N \sum_t^T \|\mathcal{T}'^t_{\text{neuron},j}(\boldsymbol{z}) - \mathcal{T}'^t_{\text{neuron},j}(\boldsymbol{z}^{\mathcal{D}})\|^2 , \tag{6}$$

where T is the number of time points in a trial and N is the number of neurons. For the real dataset, where we want to fit also the jaw movement, we have an additional term for the trial-averaged filtered and normalized jaw, $\| \sum_t^T \mathcal{T}'^t_{\text{neuron}}(\boldsymbol{y}) - \mathcal{T}'^t_{\text{neuron}}(\boldsymbol{y}^{\mathcal{D}})\|^2$, where y is the simulated jaw movement and $y^{\mathcal{D}}$ the recorded jaw movement.

(2) To calculate the **trial-matching loss**, we first filter the population-average activity of each area A, $\mathcal{T}^t_{A,k}(\boldsymbol{z}) = \frac{1}{|A|}\sum_{j\in A} \boldsymbol{z}^t_{j,k} * f$, using a rolling average window of 32 ms. We then normalize it by the population-averaged filtered activity of the same area from the recordings, $\mathcal{T}'^t_{A,k}(\boldsymbol{z}) = (\mathcal{T}^t_{A,k}(\boldsymbol{z}) - \langle \mathcal{T}^t_{A,k}(\boldsymbol{z}^{\mathcal{D}})\rangle_k)/\sigma_k(\mathcal{T}^t_{A,k}(\boldsymbol{z}^{\mathcal{D}}))$, and concatenate all the areas that were simultaneously recorded, $\mathcal{T}'^t_{\text{trial},k}(\boldsymbol{z}) = (\mathcal{T}'^t_{A1,k}, \mathcal{T}'^t_{A2,k})$, where $\langle . \rangle_k$ is the trial average, and σ_k the standard deviation over trials. The trial-matching loss is defined as:

$$\mathcal{L}_{\text{trial}} = \min_{\pi} \sum_k^K \sum_t^T \|\mathcal{T}'^t_{\text{trial},k}(\boldsymbol{z}) - \mathcal{T}'^t_{\text{trial},\pi(k)}(\boldsymbol{z}^{\mathcal{D}})\|^2 , \tag{7}$$

where π is an assignment between pairs of K recorded and generated trials $\pi : \{1, \dots K\} \to \{1, \dots K\}$. Note that the minimum over π is a combinatorial optimization that needs to be calculated for every evaluation of the loss function. For the real dataset, we consider the jaw movement as an additional area, and we concatenate it to the $\mathcal{T}'^{t}_{\text{trial},k} = (\mathcal{T}'^{t}_{A1,k}, \mathcal{T}'^{t}_{A2,k}, \mathcal{T}'^{t}_{jaw,k})$.

Based on this loss function, we optimize the following parameters: $W^{rec,d}_{ij}$, $W^{in,d}_{ij}$, $v_{thr,j}$, and β for the RefCircs. For the RNNs, we optimize only the recurrent connectivity $W^{rec,d}_{ij}$, and the rest are fixed from the RefCircs. For the real dataset, additionally to the parameters optimized in the RefCircs, we also optimize the jaw's linear readout W^{jaw}_{j} and its scaling parameter b.

Implementing Dale's law and local inhibition In our network, the recurrent weights W^{rec} are computed as the elementwise product of two matrices: $\tilde{W}^{rec}$, which encodes the strength of synaptic efficacies and is always positive, and W^{rec}_{sign}, which has a fixed sign determined by the neurotransmitter type of the presynaptic neuron and $|W^{rec}_{sign}| = 1$:

$$W^{rec} = \tilde{W}^{rec} \circ W^{rec}_{sign} \tag{8}$$

To enforce Dale's law during optimization, we set any negative values of $\tilde{W}^{rec}$ to zero at each iteration as in Bellec et al. (2018a). Similarly, to constrain only local inhibitory connections during optimization, we zero out any changes in the synaptic efficacies of across-areas inhibitory connections at each iteration. In simplified models, Dale's law or the local inhibition constraint can be disrupted by omitting this correction step.

The success of the network optimization highly depends on the initialization of the recurrent weight matrices. To initialize signed matrices we follow the theoretical Rajan and Abbott (2006) and practical insights Bellec et al. (2018b); Cornford et al. (2020) developed previously. After defining the constraints on the weight signs W^{rec}_{sign}, the initialization amplitude $\tilde{W}^{rec}$ for each target neuron is adjusted to a zero-summed input weights (the sum of incoming excitatory inputs is equal to the sum of inhibitory inputs). Then the weight amplitude is re-normalized by the modulus of its largest eigenvalue of W^{rec}, so all the eigenvalues of this matrix W^{rec} have modulus 1 or smaller.

Stopping criterion of the optimization For the synthetic dataset, we use early stopping, meaning that our final model has the smallest test loss. For the real dataset, due to limited data and a noisy test set, we select the final model based on the optimization step that yields the best trial type accuracy, derived from the jaw trace and whisker stimulus, along with the highest trial-matched Pearson correlation between the model and the recordings.

Sparsity regularization There is a plethora of ways to enforce sparsity. In this work, we use weight regularization. In particular, we use the $l_{\frac{1}{2}}$ norm of the recurrent and input weights that promote a high level of sparsity (Xu et al., 2012). To avoid numerical instabilities, we apply this regularization only for synaptic weights above α and prune all synapses below α. (we set $\alpha = 1e^{-7}$). The regularized loss function becomes:

$$\mathcal{L}_{all} = \mathcal{L} + \lambda_1 ||W^{in}||^{\frac{1}{2}}_{\frac{1}{2}} + \lambda_2 ||W^{rec,d}||^{\frac{1}{2}}_{\frac{1}{2}} + \lambda_3 ||W^{\text{across},d}||^{\frac{1}{2}}_{\frac{1}{2}}, \tag{9}$$

where $W^{\text{across},d}$ are the connections from one area to the other.

For the synthetic dataset, we choose the level of across-area sparsity by performing a small grid search for λ_3. In particular, the sparsity level λ_3 is the maximum value λ_3 where the performance remains as good as without sparsity, see Suppl. Fig S3. For the real dataset, we use the same value λ_3 as the one we found for the full reconstruction method of bioRNN1.

4.6 Perturbation test of in silico optogenetics

In systems neuroscience, a method to test causal interactions between brain regions uses spatially and temporally precise optogenetic activations or inactivations (Esmaeili et al., 2021; Guo et al., 2014). Usually, inactivations refer to the strong activation of inhibitory neurons for cortical areas. These inhibitory neurons have strong intra-area connections that effectively "silence" their local-neighborhood (Helmstaedter et al., 2009).

Our model can simulate these perturbations and allow us to compare the causal mechanisms of two networks based on their responses to optogenetic perturbations. We implement activations and inactivations as a strong input current to all the neurons in one area's excitatory or inhibitory population. For the RefCircs and bioRNNs, we use a transient current that lasts 40 ms, from 20 ms before to 20 ms after the input stimulus. The strength of the current (light power) varies until there is an effect in the full reconstruction method bioRNN1. For the real dataset, we perform optogenetics inactivations in three different periods. As in Esmaeili et al. (2021), we silence the cortical circuit during the whisker presentation, the time between the whisker and auditory stimulus, and when the animal was licking for the reward. In particular, we use transient currents to the inhibitory neurons during (i.) 100 ms before and after the whisker presentation, (ii.) 100 ms after the whisker presentation till 100ms before the onset of the auditory cue, and (iii.) after the auditory cue till the end of our simulation. For cases (i.) and (ii.), we linearly decreased the strength of the current to avoid rebound excitation. The light power is chosen so that our model has the best results in reproducing the lick probability of the recordings. It is important to mention that the perturbation data are not used to optimize the network but to test whether the resulting network has the same causal interactions with the recordings.

For the RefCircs and bioRNNs, we evaluate the effect of the perturbations directly from the neural activity. We use the distance of network dynamics $\mathcal{L}_{\text{trial}}$ to compare the two perturbed networks. For the real dataset, we compare the effect of the inactivations on the behavior; as behavior here, we mean whether the mouse/model licked. We classify the licking action using a multilayer perception with two hidden layers with 128 neurons each. The classifier is trained with the jaw movement of the real dataset, which was extracted from video filming using custom software, to predict the lick action, which was extracted from a piezo sensor placed in the spout. This classifier predicted lick correctly 94% of the time. We then used the same classifier on the jaw movement from the model to determine whether there was a "lick" or not.

References

Aravanis, A. M., Wang, L.-P., Zhang, F., Meltzer, L. A., Mogri, M. Z., Schneider, M. B., and Deisseroth, K. (2007). An optical neural interface: in vivo control of rodent motor cortex with integrated fiberoptic and optogenetic technology. *Journal of neural engineering*, 4(3):S143.

Arthur, B. J., Kim, C. M., Chen, S., Preibisch, S., and Darshan, R. (2023). A scalable implementation of the recursive least-squares algorithm for training spiking neural networks. *Frontiers in Neuroinformatics*, 17:1099510.

Azabou, M., Arora, V., Ganesh, V., Mao, X., Nachimuthu, S., Mendelson, M., Richards, B., Perich, M., Lajoie, G., and Dyer, E. (2024). A unified, scalable framework for neural population decoding. *Advances in Neural Information Processing Systems*, 36.

Bellec, G., Kappel, D., Maass, W., and Legenstein, R. (2018a). Deep rewiring: Training very sparse deep networks. *ICLR*.

Bellec, G., Salaj, D., Subramoney, A., Legenstein, R., and Maass, W. (2018b). Long short-term memory and learning-to-learn in networks of spiking neurons. *Advances in neural information processing systems*, 31.

Bellec, G., Wang, S., Modirshanechi, A., Brea, J., and Gerstner, W. (2021). Fitting summary statistics of neural data with a differentiable spiking network simulator. *Advances in Neural Information Processing Systems*, 34.

Billeh, Y. N., Cai, B., Gratiy, S. L., Dai, K., Iyer, R., Gouwens, N. W., Abbasi-Asl, R., Jia, X., Siegle, J. H., Olsen, S. R., et al. (2020). Systematic integration of structural and functional data into multi-scale models of mouse primary visual cortex. *Neuron*, 106(3):388–403.

Cornford, J., Kalajdzievski, D., Leite, M., Lamarquette, A., Kullmann, D. M., and Richards, B. (2020). Learning to live with dale's principle: Anns with separate excitatory and inhibitory units. *bioRxiv*, pages 2020–11.

Cowley, B. R., Calhoun, A. J., Rangarajan, N., Ireland, E., Turner, M. H., Pillow, J. W., and Murthy, M. (2024). Mapping model units to visual neurons reveals population code for social behaviour. *Nature*, 629(8014):1100–1108.

Défossez, A., Copet, J., Synnaeve, G., and Adi, Y. (2023). High fidelity neural audio compression. *Transactions on Machine Learning Research*.

Deny, S., Ferrari, U., Mace, E., Yger, P., Caplette, R., Picaud, S., Tkačik, G., and Marre, O. (2017). Multiplexed computations in retinal ganglion cells of a single type. *Nature communications*, 8(1):1964.

Dinc, F., Shai, A., Schnitzer, M., and Tanaka, H. (2023). Cornn: Convex optimization of recurrent neural networks for rapid inference of neural dynamics. *Advances in Neural Information Processing Systems*, 36.

Dubreuil, A., Valente, A., Beiran, M., Mastrogiuseppe, F., and Ostojic, S. (2022). The role of population structure in computations through neural dynamics. *Nature neuroscience*, 25(6):783–794.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, 14(2):179–211.

Esmaeili, V., Tamura, K., Muscinelli, S. P., Modirshanechi, A., Boscaglia, M., Lee, A. B., Oryshchuk, A., Foustoukos, G., Liu, Y., Crochet, S., et al. (2021). Rapid suppression and sustained activation of distinct cortical regions for a delayed sensory-triggered motor response. *Neuron*, 109(13):2183–2201.

Guo, Z. V., Li, N., Huber, D., Ophir, E., Gutnisky, D., Ting, J. T., Feng, G., and Svoboda, K. (2014). Flow of cortical activity underlying a tactile decision in mice. *Neuron*, 81(1):179–194.

Harris, J. A., Mihalas, S., Hirokawa, K. E., Whitesell, J. D., Choi, H., Bernard, A., Bohn, P., Caldejon, S., Casal, L., Cho, A., et al. (2019). Hierarchical organization of cortical and thalamic connectivity. *Nature*, 575(7781):195–202.

Helmstaedter, M., Sakmann, B., and Feldmeyer, D. (2009). L2/3 interneuron groups defined by multiparameter analysis of axonal projection, dendritic geometry, and electrical excitability. *Cerebral Cortex*, 19(4):951–962.

Hodgkin, A. L. (1958). The croonian lecture-ionic movements and electrical activity in giant nerve fibres. *Proceedings of the Royal Society of London. Series B-Biological Sciences*, 148(930):1–37.

International Brain Laboratory, Benson, B., Benson, J., Birman, D., Bonacchi, N., Carandini, M., Catarino, J. A., Chapuis, G. A., Churchland, A. K., Dan, Y., et al. (2023). A brain-wide map of neural activity during complex behaviour. *biorxiv*, pages 2023–07.

Isbister, J. B., Ecker, A., Pokorny, C., Bolaños-Puchet, S., Santander, D. E., Arnaudon, A., Awile, O., Barros-Zulaica, N., Alonso, J. B., Boci, E., et al. (2023). Modeling and simulation of neocortical micro-and mesocircuitry. part ii: Physiology and experimentation. *bioRxiv*, pages 2023–05.

Kim, C. M., Finkelstein, A., Chow, C. C., Svoboda, K., and Darshan, R. (2023). Distributing task-related neural activity across a cortical network through task-independent connections. *Nature Communications*, 14(1):2851.

Lappalainen, J. K., Tschopp, F. D., Prakhya, S., McGill, M., Nern, A., Shinomiya, K., Takemura, S.-y., Gruntman, E., Macke, J. H., and Turaga, S. C. (2023). Connectome-constrained deep mechanistic networks predict neural responses across the fly visual system at single-neuron resolution. *bioRxiv*, pages 2023–03.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

Liu, Y., Foustoukos, G., Crochet, S., and Petersen, C. C. (2022). Axonal and dendritic morphology of excitatory neurons in layer 2/3 mouse barrel cortex imaged through whole-brain two-photon tomography and registered to a digital brain atlas. *Frontiers in neuroanatomy*, 15:791015.

Mahuas, G., Isacchini, G., Marre, O., Ferrari, U., and Mora, T. (2020). A new inference approach for training shallow and deep generalized linear models of noisy interacting neurons. *Advances in Neural Information Processing Systems*, 33.

Markram, H., Muller, E., Ramaswamy, S., Reimann, M. W., Abdellah, M., Sanchez, C. A., Ailamaki, A., Alonso-Nanclares, L., Antille, N., Arsever, S., et al. (2015). Reconstruction and simulation of neocortical microcircuitry. *Cell*, 163(2):456–492.

Markram, H., Toledo-Rodriguez, M., Wang, Y., Gupta, A., Silberberg, G., and Wu, C. (2004). Interneurons of the neocortical inhibitory system. *Nature reviews neuroscience*, 5(10):793–807.

Neftci, E. O., Mostafa, H., and Zenke, F. (2019). Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Processing Magazine*, 36(6):51–63.

Pals, M., Sağtekin, A. E., Pei, F., Gloeckler, M., and Macke, J. H. (2024). Inferring stochastic low-rank recurrent neural networks from neural data. *arXiv:2406.16749*.

Pandarinath, C., O'Shea, D. J., Collins, J., Jozefowicz, R., Stavisky, S. D., Kao, J. C., Trautmann, E. M., Kaufman, M. T., Ryu, S. I., Hochberg, L. R., et al. (2018). Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature Methods*, 15(10):805–815.

Perich, M. G., Arlt, C., Soares, S., Young, M. E., Mosher, C. P., Minxha, J., Carter, E., Rutishauser, U., Rudebeck, P. H., Harvey, C. D., et al. (2020). Inferring brain-wide interactions using data-constrained recurrent neural network models. *bioRxiv*, pages 2020–12.

Pillow, J. W., Shlens, J., Paninski, L., Sher, A., Litke, A. M., Chichilnisky, E., and Simoncelli, E. P. (2008). Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995–999.

Pozzorini, C., Mensi, S., Hagens, O., Naud, R., Koch, C., and Gerstner, W. (2015). Automated high-throughput characterization of single neurons by means of simplified spiking models. *PLoS Computational Biology*, 11(6):e1004275.

Rajan, K. and Abbott, L. F. (2006). Eigenvalue spectra of random matrices for neural networks. *Physical review letters*, 97(18):188104.

Richards, B. A. and Kording, K. P. (2023). The study of plasticity has always been about gradients. *The Journal of Physiology*, 601(15):3141–3149.

Rosenblatt, F. (1960). Perceptual generalization over transformation groups. *Self Organizing Systems*, pages 63–96.

Schrimpf, M., McGrath, P., Margalit, E., and DiCarlo, J. J. (2024). Do topographic deep ann models of the primate ventral stream predict the perceptual effects of direct it cortical interventions? *bioRxiv*, pages 2024–01.

Siegle, J. H., Jia, X., Durand, S., Gale, S., Bennett, C., Graddis, N., Heller, G., Ramirez, T. K., Choi, H., Luviano, J. A., et al. (2021). Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature*, 592(7852):86–92.

Sourmpis, C., Petersen, C., Gerstner, W., and Bellec, G. (2023). Trial matching: capturing variability with data-constrained spiking neural networks. *Advances in Neural Information Processing Systems*, 36.

Staiger, J. F. and Petersen, C. C. (2021). Neuronal circuits in barrel cortex for whisker sensory perception. *Physiological reviews*, 101(1):353–415.

Steinmetz, N. A., Zatka-Haas, P., Carandini, M., and Harris, K. D. (2019). Distributed coding of choice, action and engagement across the mouse brain. *Nature*, 576(7786):266–273.

Tamamaki, N. and Tomioka, R. (2010). Long-range gabaergic connections distributed throughout the neocortex and their possible function. *Frontiers in neuroscience*, 4:202.

Teeter, C., Iyer, R., Menon, V., Gouwens, N., Feng, D., Berg, J., Szafer, A., Cain, N., Zeng, H., Hawrylycz, M., et al. (2018). Generalized leaky integrate-and-fire models classify multiple neuron types. *Nature Communications*, 9(1):709.

Udvary, D., Harth, P., Macke, J. H., Hege, H.-C., de Kock, C. P., Sakmann, B., and Oberlaender, M. (2022). The impact of neuron morphology on cortical network architecture. *Cell Reports*, 39(2).

Urai, A. E., Doiron, B., Leifer, A. M., and Churchland, A. K. (2022). Large-scale neural recordings call for new insights to link brain and behavior. *Nature Neuroscience*, 25(1):11–19.

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

Xu, Z., Chang, X., Xu, F., and Zhang, H. (2012). $l_\{1/2\}$ regularization: A thresholding representation theory and a fast solver. *IEEE Transactions on neural networks and learning systems*, 23(7):1013–1027.

Method name	RefCirc1 vs RNN1		RefCirc2 vs RNN2	
	no light	light	no light	light
bioRNN	0.19 ± 0.01	0.21 ± 0.05	0.18 ± 0.01	0.20 ± 0.08
σRNN	0.16 ± 0.01	0.47 ± 0.64	0.17 ± 0.01	0.56 ± 0.53
No sparsity	0.20 ± 0.01	0.48 ± 0.79	0.19 ± 0.01	0.19 ± 0.06
Non-local inhibition	0.20 ± 0.02	0.37 ± 0.31	0.18 ± 0.01	0.68 ± 0.69
No Dale's law	0.18 ± 0.01	0.41 ± 0.31	0.18 ± 0.01	0.84 ± 1.11
No TM	0.33 ± 0.01	0.38 ± 0.10	0.35 ± 0.03	0.37 ± 0.09
No spike	0.17 ± 0.00	0.18 ± 0.02	0.18 ± 0.00	0.26 ± 0.15

Table 1: The loss $\mathcal{L}_{\text{trial}}$ measures the distance between the network dynamics in single trials. Here for the synthetic dataset, we report the distance between activity statistics in area A during stimulation of area B. The column "light" reports the distance between RNN and RefCirc dynamics during perturbation, the column "no light" reports the same value when no stimulation is delivered. ± indicates the 95% confidence interval.

Method name	Real Dataset vs Reconstructed network
bioRNN	0.76 ± 0.14
σRNN	0.62 ± 0.12
No sparsity	0.77 ± 0.15
Non-local inhibition	0.79 ± 0.15
No Dale's law	0.68 ± 0.13
No TM	1.63 ± 0.55
No spike	0.64 ± 0.13

Table 2: Trial-matching loss test loss $\mathcal{L}_{\text{trial}}$ of the different reconstruction methods with the real recordings from (Esmaeili et al., 2021) ± indicates the 95% confidence interval.

Supplemental information

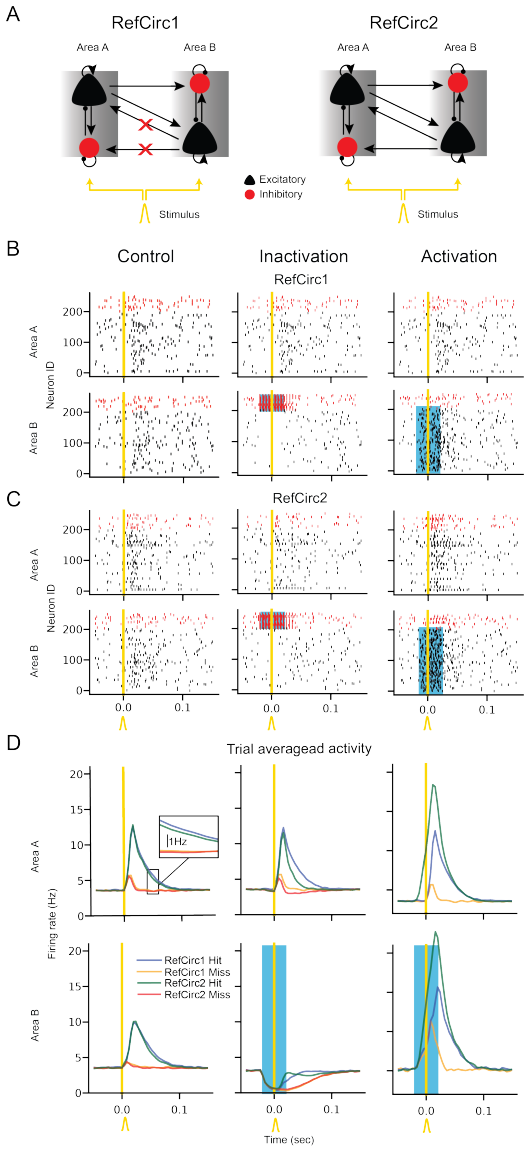

Figure S1: **Modeling "optogenetic" perturbations. A.** Two different network hypotheses for implementing a detection task. In RefCirc1, area A projects to area B but not vice versa. In RefCirc2, the areas are recurrently connected. **B.** Raster plots of all neurons in RefCirc1 during a single hit trial under normal conditions (control, left) and under optogenetic perturbation of excitatory (middle) and inhibitory (right) neurons. The duration of the light stimulus is shown with a blue shading. **C.** Same for RefCirc2 **D.** Trial-averaged activity of the two circuits during Hit (blue: RefCirc1; green: RefCirc2) and Miss (yellow: RefCirc 1; red: RefCirc2) trials. A trial is classified as "Hit" if both areas reach a transient firing rate above 8Hz; and otherwise as "Miss". Neurons directly targeted by the light are excluded from the averaging. For the control case, the maximal difference between the trial average activity of the two networks is below 0.51 Hz (zoom inset).

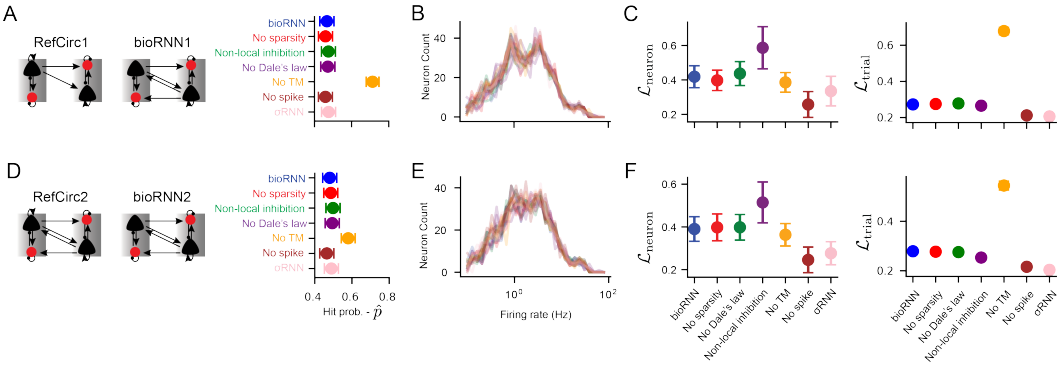

Figure S2: **Fitting Reconstructed networks to the synthetic dataset. A.** Schematic representation of the RefCirc1 and bioRNN1. and probability of hit trials. **B.** Histogram of the firing rate distribution of the RefCirc1 and all the RNN1 versions. We observe that all RNN1 versions fit well with the RefCirc1. **C.** Left: Neuron loss of the different RNN1 variants. Right: Trial-matching loss of the different RNN1 variants. We observe that the model without the trial-matching loss function behaves considerably worse. The whiskers show the 95% confidence interval of the mean across trials. **D-F.** Same as **A-B** for RefCirc2 and RNNs2.

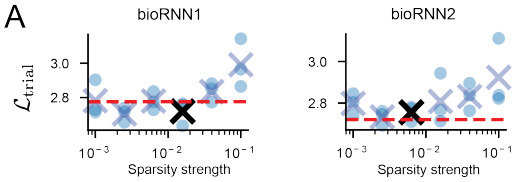

Figure S3: **Picking the sparsity level. A.** Grid search for the optimal maximum regularization strength (λ_3) without a drop in performance. As a performance measure, we used the trial-matching loss, $\mathcal{L}_{\text{trial}}$.

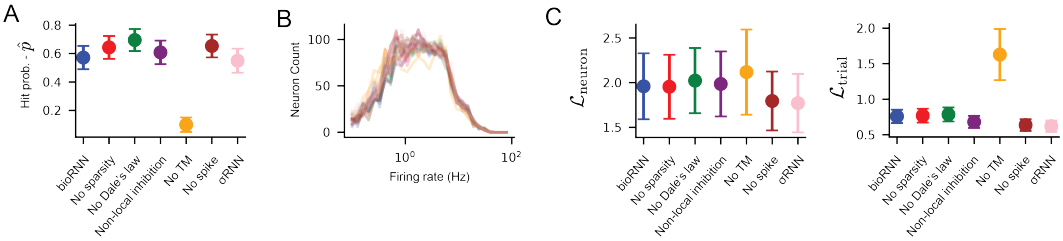

Figure S4: **Reconstruction of the real recordings. A.** Probability of hit trials of the different variant models. **B.** Histogram of the firing rate distribution from the real recordings and all the variants. **C.** Top: Neuron loss of the different RNN1 variants. All RNN versions have a similar loss value. Bottom: Trial-matching loss of the different model variants. We observe that the model without the trial-matching loss function behaves considerably worse. The whiskers show the 95% confidence interval.

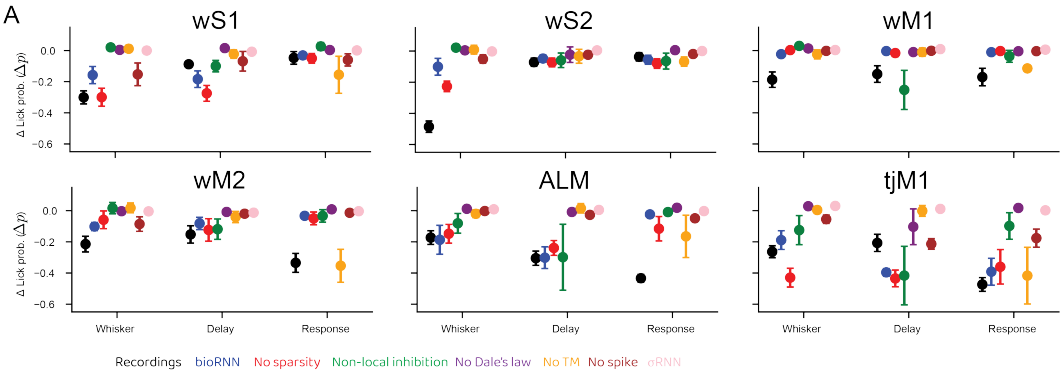

Figure S5: **A** Change of lick probability under inactivation of all areas in all the different temporal windows. We show the Δp from the data and reconstruction model variants.