Biologically informed cortical models predict optogenetic perturbations

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

Abstract

Ideally a recurrent spiking neural network fitted to large electrophysiological data sets should illustrate the chain of cortical information transmission. In particular, successful network model reconstruction methods should enable a model to predict the response to optogenetic perturbations. We propose a method that extracts a recurrent spiking neural network from electrophysiological data sets and test it on unseen optogenetic interventions. Our method uses multiple objectives to reproduce various aspects of neuronal activity patterns as well as biology-based inductive biases like structured connectivity and separation of neurons into dominant excitatory and inhibitory cell types. We show that theses biological inductive biases are crucial to predict responses to optogenetic perturbations. Furthermore, once a model is fitted to experimental data, gradient evaluations in the extracted model network can, in principle, be used to design a minimal low-power optogenetic stimulus that causes a maximal effect by stimulating a limited subset of neurons. Our computational methods provide a systematic path towards translating observed spike data into a matching recurrent neural network that is able to predict the results of a variety of experimental interventions.

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). Often, a rodent is trained to perform a simple task involving sensation and a rewarded behavioral action. The animal's behavior is tracked with cameras, and the neural activity is recorded simultaneously across multiple cortical areas. With an increasing density of electrodes and a coverage of the entire sensory-motor pathway, this experimental approach promises to improve our understanding of the complex mechanisms taking place. However, the field lacks data-grounded modeling approaches to generate and test hypotheses on the causal role of neuronal and circuit-level mechanisms.

There is a long tradition of data-informed neuroscience simulations. Biophysical models have been crucial for our understanding of single-cell mechanisms (Hodgkin, 1958), and have more recently been used to describe neuronal interactions across cortical layers, columns, and areas (Markram et al., 2015; Billeh et al., 2020; Isbister et al., 2023). Classically, these data-grounded simulations select a subset of established knowledge from the neuroscience literature (such as neuronal time constants, connectivity statistics, cell morphology) to constrain the simulation (Markram et al., 2015; Billeh et al., 2020; Isbister et al., 2023; Reimann et al., 2024). However, fine-tuning a large simulation to include and match new data points is labor- and compute-intensive. To leverage the high information density of contemporary recordings, new modeling approaches are needed.

A promising approach resides in the optimization of the simulation parameters by gradient descent. These methods were successful in quantitatively describing and classifying

functional cell types (Pozzorini et al., 2015; Teeter et al., 2018), and modeling micro-circuit interactions (for successful applications to the retina see 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 works made substantial progress towards data-constrained recurrent neural networks (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). This line of work leverages modern optimization methods from machine learning to predict recorded activity at large. Nevertheless, such an approach is different from generic deep learning methods (Schrimpf et al., 2018; Pandarinath et al., 2018) as the one-to-one mapping between a recorded and simulated neurons provide a mechanical interpretation of individual cells in the networks Cowley et al. (2024).

In this context, an important question is whether these data-constrained model networks are indeed capable of inferring mechanisms connecting individual neurons with macroscopic cortical activity and behavior. By construction, these networks are expected to explain the recorded electrophysiological, and potentially anatomical, data, but how can we measure whether the causal mechanism in the reconstructed network faithfully represents biophysical reality? To answer this question, we submit a range of circuit reconstruction methods to a difficult *perturbation test*: Can the reconstructed network produce the same response to causal perturbation as the recorded biological networks?

Optogenetics is a powerful tool to induce precise causal perturbations in vivo (Esmaeili et al., 2021; Guo et al., 2014). This technique involves the expression of light-sensitive ion channels (Aravanis et al., 2007), such as channelrhodopsins, in specific populations of neurons (e.g., inhibitory or excitatory neurons). The cation channels, activated upon illumination, depolarize and activate the corresponding cell type in a selected cortical area. Contemporary dense electrophysiological recordings combined with optogenetic perturbations are an appealing benchmark to compare network reconstruction methods (Esmaeili et al., 2021). In this paper, we develop a range of network reconstruction methods that suit the datasets at hand: (1) a simplified spiking neuron model enables a direct mapping to spike-sorted electrophysiology recordings, and (2) the simulated neurons can be mapped to the dominant cell-types, i.e., excitatory/pyramidal and inhibitory/parvalbumin-expressing (PV), which are selectively targeted in optogenetic perturbations. Parameters of single neurons as well as synaptic connection strengths are optimized to build a data-constrained recurrent spiking neural network that explains the recorded spiking activity.

After optimization, we can test the reconstruction method by comparing the response to optogenetic perturbations in the recorded circuit and the reconstructed recurrent neural network (ReconRNN). We report which aspects of the reconstruction method have the strongest impact on this *perturbation test*. Interestingly, the robustness under causal perturbations varies greatly, even for methods that successfully fit the electrophysiology recordings. We find that adding biological assumptions in the ReconRNN (so-called biological inductive biases) improves the robustness of the network to simulated optogenetic perturbations.. Namely, two constraints on the network topology have a strong effect: (1) Dale's law (outgoing synapse transmission efficacy is positive or negative according to the cell type), and (2) local-only inhibition (inhibitory neurons do not project to other cortical areas). These two constraints are important for network reconstruction both on a synthetic dataset and a real electrophysiology dataset collected in mice in vivo (Esmaeili et al., 2021). On the synthetic dataset, we find that (3) a sparsity prior on the network connectivity can also improve the quantitative prediction.

We further show an equivalence between the measured effects to small perturbations and gradients in the recorded dynamical system. This theoretical analysis is confirmed in silico. Considering the foundational role of gradients in machine learning theory, we argue that network reconstruction methods robust to perturbations opens the way for a computational understanding of cortical recordings.

2 Results

2.1 Synthetic dataset challenging for causal inference

The goal is to understand the chain of causal interactions between the relevant cortical areas that drive behavior. Datasets that capture activity throughout a sensory-motor pathway can involve six areas or more (Steinmetz et al., 2019; Esmaeili et al., 2021; International Brain Laboratory et al., 2023), but we first identify a common sub-problem: let's consider two areas (Area A and B), and we ask whether the areas are connected in an exclusively feedforward fashion (hypothesis $H1$) or whether there is a strong recurrent interaction (hypothesis $H2$). This common interrogation is relevant between a primary and a secondary sensory cortical area or a relevant pair of sensory-motor or motor-motor areas. We build a simpler synthetic dataset capturing this difficulty. This dataset will be the first benchmark to illustrate and analyze quantitatively our reconstruction methods.

The synthetic data set has two areas A and B which are either transiently active together ("Hit trial" occurring with frequency p_{Hit}) or quiescent together ("Miss trial" occurring with probability $1 - p_{\text{Hit}}$). Since the macroscopic activity does not display a clear hierarchy between A and B it is not trivial to distinguish the feedforward ($H1$) from the recurrent hypothesis ($H2$). We show mathematically in Methods 4.1 that both a feedforward and a recurrent statistical model generate the same probabilistic outcome, making it necessary to investigate dense neural activity recording in more details to decide between hypotheses $H1$ and $H2$ without increasing the number of different stimuli or adding perturbations. Such a difficult case is encountered regularly in real datasets because of the binary (or low dimensional) nature of the task stimuli and behavioral outcomes.

To generate artificial data sets of spike train recordings, we switch from a macroscopic statistical model to two reference circuits, RefCirc1 and RefCirc2, corresponding to the feedforward ($H1$) and recurrent hypotheses ($H2$), respectively. The two circuits not only display a bi-modal response with the same p_{Hit} probability, but also very similar neural activity. 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 inhibitory. In both reference circuits, only excitatory neurons have connections to the other area. Triggered by a short input pulse, areas A and B of both RefCircs are near-simultaneously active ("Hit trials") or inactive ("Miss trials") in approximately half of the trials, see Figure 1. A trial is considered a "Hit" when the maximum firing rate of the population average in both areas is above 7 Hz. By construction of the synthetic dataset, each neuron in RefCirc1 has the identical trial-averaged activity as the corresponding neuron in RefCirc2. Moreover, the population-averaged activity shows nearly the same activity profile during Hit trials for the two RefCircs as shown in the Control column in Figure 1 (see Methods 4.4 for more details). Essentially, the only difference between the two RefCircs is not visible in the spiking data 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. From each of the two RefCircs, we generate a dataset by simulating 2,000 trials of spike train recordings of all 500 neurons from areas A and B simultaneously. Half of these trials are the training dataset and will be the basis for our data-driven circuit reconstruction method. The second half of the recorded trials form the testing-set which is used as a first evaluation of the quality of the reconstruction.

The difference between the feedforward RefCirc1 and the recurrent RefCirc2 becomes visible with an optogenetic perturbation of the network. 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 of an area is simulated as a positive current injected into excitatory cells). Figure 1B-D shows that optogenetic perturbations in area B reflect the presence or absence of feedback connections. We will later use these simulated optogenetic perturbations as the ultimate evaluation method to

probe if the reconstructed networks (ReconRNN1 and ReconRNN2) reflect the mechanisms of the reference circuits RefCirc1 and RefCirc2.

2.2 Reconstructed Networks: biological inductive biases strengthen robustness to perturbations

Network reconstruction methodology We optimize parameters of a recurrent neural network so as to optimally approximate the spike trains of a data set from a RefCirc. Parameter optimization ends with a final Reconstructed Recurrent Neural Network (ReconRNN). For our synthetic data set, the ReconRNN has the same size and the same number of excitatory and inhibitory neurons as the RefCirc, but neuronal parameters as well as network connectivity are unknown. Importantly, at initialization, the networks are blind to the connectivity structure of the RefCirc (feedforward or recurrent). In particular we allow all excitatory neurons to project to any neuron in the network with transmission efficacies (aka as synaptic weights) initialized randomly. To turn this initial network into a ReconRNN, we optimize the synaptics "weights" to minimize the difference between the reconstructed and reference activity statistics. The optimization of these parameters minimizes three loss functions: (i) Each neuron in the dataset is mapped to a single neuron in the ReconRNN, and the neuron-specific loss function $\mathcal{L}_{\text{neuron}}$ is the mean-square error of the *trial-averaged* spike counts of the ReconRNN and RefCirc neurons (aka. peristimulus time histogram, PSTH). (ii) To minimize the distance between single-trial network activity, we follow the methodology from Sourmpis et al. (2023) and use a trial-specific loss function $\mathcal{L}_{\text{trial}}$, which is the distance between the distribution of single trial population-averaged activity of the ReconRNN and RefCirc (so-called "Trial-matching" loss function). (iii) Finally, we implement the prior that biological networks are sparsely connected with a regularization loss function $\mathcal{L}_{\text{reg}}$. This regularization loss can also include the prior that local connections within an area are more dense than across areas. The three losses are defined mathematically in Methods 4.5.

To identify the effect of individual features of the reconstruction method, we train multiple ReconRNN variants which are deprived of one feature. For all these model variants and after minimization of the loss function, we verify that the network trajectories closely match the data on the testing-set in terms of (i) the "behavioral" hit-trial frequency (a hit-like trial is defined when the population average firing rate is higher than 7Hz in both areas), (ii) the peristimulus time histogram (PSTH) pearson correlation of single neurons, and (iii) the distance between single-trial network dynamics as evaluated by $\mathcal{L}_{\text{trial}}$ (see Suppl. Fig. S1).

Based on these results, we identify two types of ReconRNN variants. First, whenever the loss function term $\mathcal{L}_{\text{trial}}$ is omitted from the optimization procedure, at least one of the above test metrics is severely impaired. In contrast, whenever we omit data-inspired specifications (aka "biological inductive biases") the ReconRNNs do not have obvious deviations. Namely these biological inductive biases are: (1) the "No Dale's law" model, which does not have excitatory and inhibitory weight constraints, (2) the "Non-local inhibition" model, which allows inhibitory neurons to project outside of their areas, and the (3) "No Sparsity" model, where the loss $\mathcal{L}_{\text{reg}}$ is omitted. The accuracy metrics on the testing sets are reported for all ReconRNN variants on the Suppl. Fig. S1.

Perturbation test To test which of the ReconRNNs capture the causal mechanisms of the RefCircs, we simulate optogenetic activations and inactivations of area B. We first compare the change of hit probability after perturbations in ReconRNN ($\Delta\hat{p}$) and in RefCirc ($\Delta p^{\mathcal{D}}$) in Figure 2. In this synthetic dataset, only the ReconRNNs with all three biological inductive biases ((1) Dale's law, (2) local-inhibition and (3) sparsity) capture the behavioral response of the RefCirc during optogenetic perturbations. A model deprived from any of those three features could not match the frequency of hit trials under perturbations. To illustrate this result, we display the trial-averaged activity in area A during perturbed hit trials in Figure

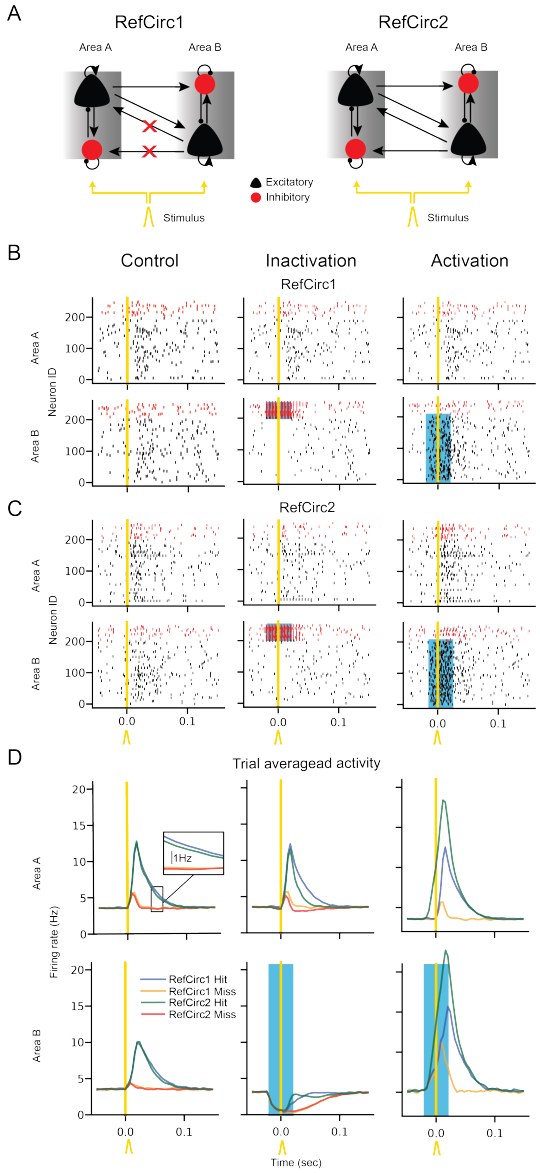

Figure 1: **Modeling "optogenetic" perturbations. A.** Two different network hypotheses for implementing a detection task. In RefCirc1, area A projects to area B but not vice versa. In RefCirc2, the areas are recurrently connected. **B.** Raster plots of all neurons in RefCirc1 during a single hit trial 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 7Hz; 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).

2B and report the distance to the reference in terms of single-trial network dynamics in Figure 2D.

We first analyze the ReconRNN variant having strictly no sign-constraints (No Dale's law). This variant is close to the data-constrained RNN fitted on electrophysiology data in previous studies Pillow et al. (2008); Bellec et al. (2021); Perich et al. (2020); Arthur et al. (2023); Kim et al. (2023). Consistently with previous work on sign-constrained optimization Bellec et al. (2018a); Cornford et al. (2020) we find that the Full reconstruction method and the un-constrained model (No Dale Law's) achieve similar test-accuracies (the average distance to recorded trial dynamics are not significantly different: p-values are 0.38 or 0.13 for RefCirc1 and RefCirc2 respectively). Despite that, this unconstrained model is less robust to optogenetic perturbation in Figure 2: the hit-rate under perturbation of area B in the feedforward reconstruction ReconRNN1 is largely over estimated (Figure 2C) and the perturbed firing rate in area A is distant from the reference (purple curve in Figure 2B). These results suggest that our simulated optogenetic perturbation test is a hard case of out-of-distribution generalization.

The intermediate model (non-local inhibition) with partial sign constraints on the weight matrix can help us to understand this failure. Under inactivation of area B, this model exhibits a reduction in the baseline firing rate of area A (green trace during inactivation in Figure 2B). Intuitively this is expected because by design, in this model, inhibitory neurons from area B have mono-synaptic inhibitory effect onto area A. It is not clear however that this effect cannot be compensated by the optimization method because the most robust parameter solution (Full reconstruction method) is included in the search space of the unconstrained model. In principle, this optimization should be able to retrieve this solution or compensate the effect of long-range inhibitory connections with other mechanisms. But instead, the optimization found equally-optimal solutions that did not exhibit the robust properties of the biologically informed parameter constraint. Altogether, the two model variants with weaker constraints (No Dale's law and non-local inhibition) illustrate why adding the biological inductive bias on the connectivity of inhibitory and excitatory cell types improves the robustness to optogenetic perturbations.

Going further, the constraint on the sign of the weights is, on its own, not sufficient to recover the correct connectivity of the RefCirc. Rather we also need to add the sparsity prior on the connectivity. The strength of the feedback appears to be over estimated in both the feedforward and recurrent reconstruction in the "No Sparsity" model (red curve in Figure 2B). Adding the sparsity prior resolves this issue (blue curve). On this dataset, only the model with constrained weights and sparsity prior captures the characteristic differences between feedforward and recurrent architectures RefCirc1 and RefCirc2.

In summary, this synthetic dataset illustrates well that weight and sign constraints and sparsity enables the network to correctly predict simulated optogenetic perturbations. We will now investigate whether the same approach can predict optogenetic perturbations on a real dataset recorded in vivo.

2.3 Validating causal 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 to lick in order 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

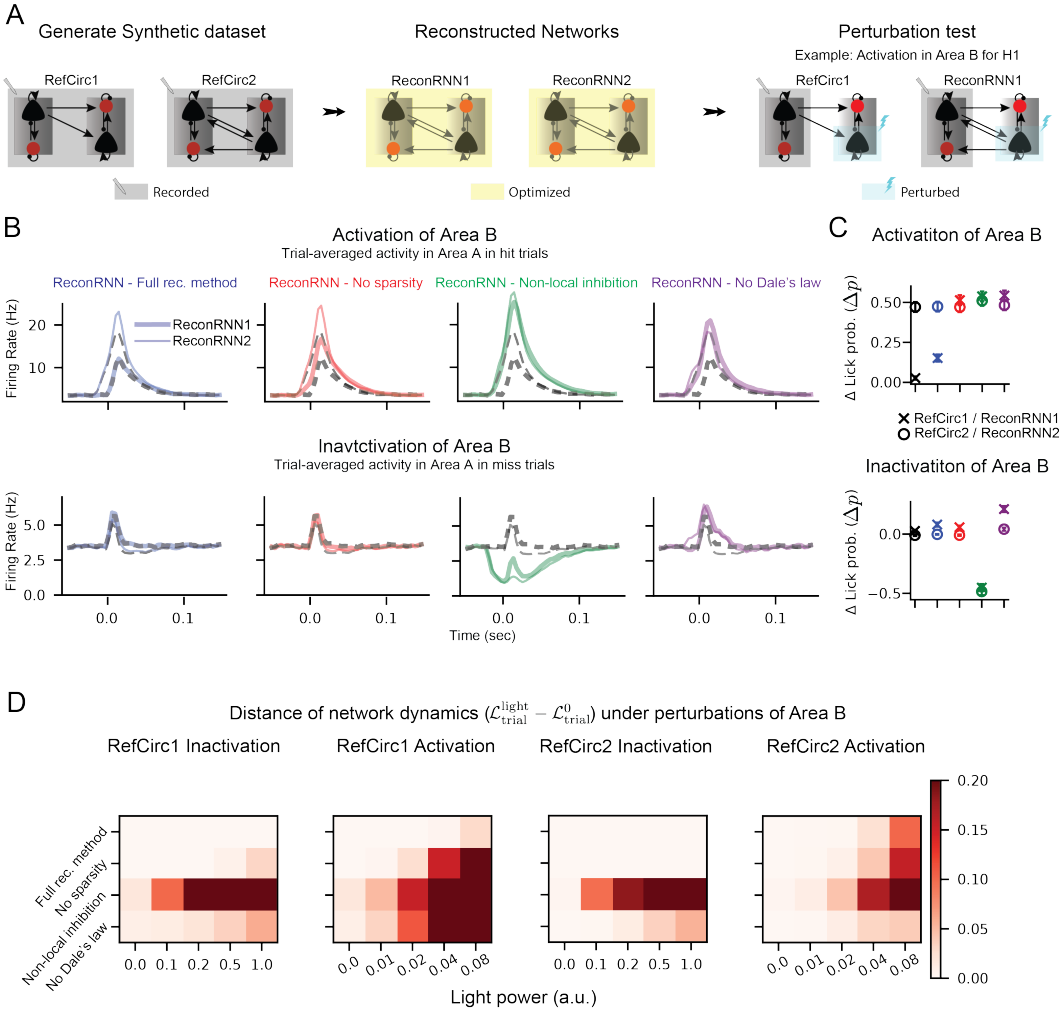

Figure 2: **Biologically informed network models predict optogenetic perturbations. A.** We optimize the parameters of the ReconRNN1 and ReconRNN2 based on spiking data from one of the RefCirc1 and RefCirc2, respectively. The ReconRNNs are blind to the structure of the RefCircs, but must extract connectivity data from the spiking data alone. After optimization, we subject the ReconRNNs to a perturbation test by applying the same simulated optogenetic perturbations as in the RefCirc. **B.** Trial-averaged activity in area A under activation/inactivation of area B. Dashed black lines indicate activity of RefCirc1 (thick dashed) and RefCirc2 (thin dashed). Results from the full reconstruction method (blue, left panel) are compared to those without the sparsity constraint (red, second panel), without local inhibition (green, third panel) and without Dale's law (magenta, right). **C.** Change in hit probability after perturbations Δp. Since we perturbed area B, a trial is classified as a "Hit" if the firing rate in area A is over 7Hz. **D.** Distance of network dynamics ($\mathcal{L}^{\text{light}}_{\text{trial}} - \mathcal{L}^{0}_{\text{trial}}$) between area A of the RefCirc and area A of the ReconRNN, as a function of perturbation strength applied to area B (horizontal axis: light power in arbitrary units). Dark colors indicate that the dynamics of the ReconRNN are different from those of the RefCirc.

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 loss function as used for the synthetic data from RefCirc1 and RefCirc2. After training, our reconstructed model can generate neural activity with similar firing rate distribution, trial-averaged activity, and single-trial network dynamics as in the recordings (see Suppl. Fig S3). Consistently with the synthetic dataset, we observe again that the model optimized without Dale's law fits the unperturbed data slightly better than does the model with Dale's law (see Table 2 and Fig. S3).

To test if the resulting network has the same mechanisms as the sensory-to-motor pathway, we subject our model to a *perturbation test*. These perturbations are acute spatiotemporal optogenetic inactivations of each area during different time periods (see Figure 3B). In the data, we measure the change in lick probability due to the optogenetic perturbations. To generate a comparable lick action in the model, 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. This optimization is performed jointly when fitting the synaptic weights to the neuronal recordings, see Methods 4.5. A binary classifier is used to identify when a jaw movement trace corresponds to a "lick" action on the water spout. In Figure 3C, we show the spiking activity of our model for a single trial where the classification changed from Hit to Miss after inactivating wS1 during the Whisker period.

We denote $\Delta p^{\mathcal{D}}$ as the change of lick trial frequency in the data during Go trials after perturbations, and $\Delta \hat{p}$ is the corresponding change probability in the model. The results of this perturbation test are reported in Figure 3E and Suppl. Fig. S4 where we compare $\Delta \hat{p}$ with $\Delta p^{\mathcal{D}}$ across brain areas, trial period (stimulation, delay or choice period) and ReconRNN model variants. On average, the error of change probability obtained with the model without sign constraints is consistently larger (see Figure 3D). So consistently with the synthetic dataset, we find that imposing Dale's law and local inhibition improves the robustness of the model to optogenetic perturbations. Conversely, adding the sparsity prior does not improve the perturbed hit-rate prediction on the real data.

To further verify the consistency of the perturbations in the model, we display in Figure 4 the perturbation map showing changes in lick propability obtained from acute inactivation in the data and the model. This perturbation map summarizes visually which area has a critical role at specific time-points. The optogenetic inactivation maps obtained in the model and data agree qualitatively on three dominant features: (1) inactivation of the sensory areas is only critical during the stimulation, (2) inactivation during the delay is most critical during the delay and (3) inactivation during the choice period is most critical in tjM1. In the model, we can easily simulate a more detailed map that would be costly to implement in vivo. On the right panel we simulated 200 trials with optogenetic perturbations for each area and 100ms time window, this corresponds approximately to 10 times more trials than what is available in the in vivo recordings Esmaeili et al. (2021). We see that the sensory areas are most critical during sensation. This visualization exhibit a first concrete prediction about the data: ALM is most critical early during the delay period, but the role switches to tjM1 during the delay.

Since the model appears to enable non-trivial causal interventions in the data, we wondered how our data-constrained modeling could be used to make computational interpretations or predictions on the data. Conceptually, we prove in the model that causal interventions enable a measurement of gradients of the recorded system. Gradients are the fundamental theoretical object to understand learning and memorization in machine learning LeCun et al. (2015), and are having a growing influence in our theoretical understanding of the brain Richards and Kording (2023). In our data-constrained model we verify the following relationship. Let's consider a behavioral measurement such as the summed jaw movement $Y = \sum_t y_t$, then we denote ΔY as the change induced by an optogenetic inter-

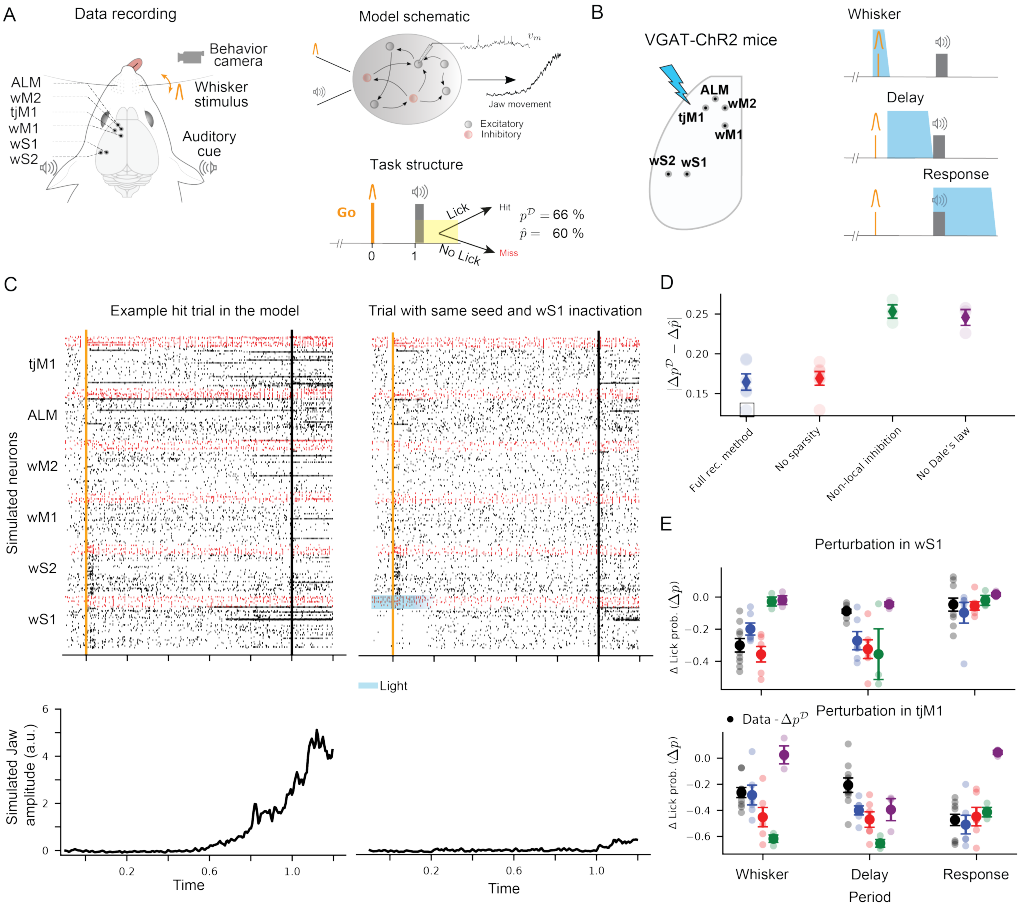

Figure 3: **Predicting optogenetic perturbations in vivo A.** During a delayed whisker detection task, the mouse reports a whisker stimulation by licking to obtain a water reward. Jaw movements are recorded by a camera. Our model simulates the neural activity in six areas as well as jaw movements of the mouse. **B.** In Esmaeili et al. (2021), experimentalists performed optogenetic spatiotemporal inactivations (blue) of several areas (one area at a time) during three different temporal windows, namely the whisker, delay, and response periods. **C.** Example hit trial of a reconstructed network (left; predicted jaw movement in the bottom graph). Simulating with the same random seed, the trial turns into a miss trial if we inactivate area wS1 during the whisker period. **D.** The average absolute difference of the change in lick probability between the different models and the recordings. Data points in light gray indicate different recording sessions for the data and different network initializations for the model, the solid points are the average performance, and the whiskers are the standard error of mean across initializations. **E.** Change of lick probability under perturbation of wS1 (Top) and tjM1 (Bottom) in the different temporal windows (see Suppl. Fig. S4 for the other areas). We show the Δp from the data and reconstruction model variants.

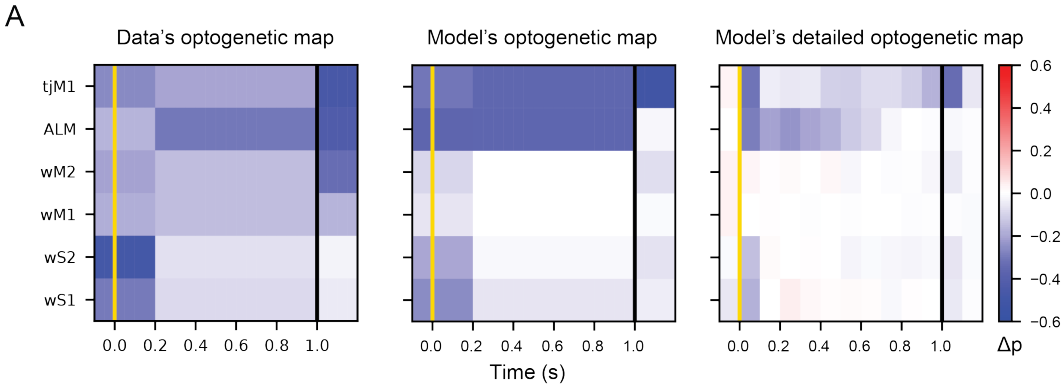

Figure 4: **optogenetic inactivation maps A.** Change of lick probability recorded with optogenetic inactivations in the experiment (left) and in a ReconRNN model (middle and right). In the model we perform the inactivation with the same light duration as used in the data (middle) or with smaller time windows (right). The optogenetic map simulated with smaller time-windows requires 10 times as many trials as available in the experiment.

vention. We perform a first order Taylor expansion to justify the following approximation in our ReconRNN model:

$$\Delta Y = \sum_{i,t \in \mathcal{I}} \frac{\partial Y}{\partial u_i^t} \Delta u_i^t, \tag{1}$$

where $\mathcal{I}$ is the set of neuron and time indices that we inject a current Δu_i^t to neuron i to simulate the optogenetic intervention. This formula tells us that a behavioral measurements under perturbations is an approximate measurement of the gradient $\frac{\partial Y}{\partial u_i^t}$ in the recorded circuit.

To verify this approximation in the model, we show that the gradient $\frac{\partial Y}{\partial u_i^t}$ can be used to target micro-perturbations (μ-perturbations) inducing a maximal behavioral change. We display in Figure 5 gradients $\sum_{i,t} \frac{\partial Y}{\partial u_i^t}$ in the form of fine-grained sensitivity map to display the effect that is predicted on the behavior. These maps are remarkably heterogeneous across trials, suggesting that the neuron-identity and time window at which the network is most-sensitive depends highly on the trial context. We use these maps to predict the outcome of μ-perturbation as follows: for each trial, and each 50ms time window, we identify 20 neurons in the model with highest gradients $\frac{dY}{du_i^t}$. We then re-simulate the exact same trial with identical random seed, but induce a μ-perturbation on the most critical neurons (see rectangles in Figure 5). If we target neurons with positive gradients, the perturbed jaw movements are strongly amplified; conversely, if we target neurons with negative gradients the jaw movements are suppressed.

3 Discussion

Selecting the relevant level of model details to capture a biophysical model is always a puzzle opening intense discussions among modelers and neuroscientists. Our results show that co-development of in vivo causal interventions with biological modeling sets a rigorous approach to quantify which model features are necessary to capture an accurate chain of causal mechanisms. *Perturbation tests* emerge as the fundamental way to set apart an efficient brain model from two classes of erroneous models which are frequent in neuroscience: (1) machine learning black-boxes with predictive power but capturing a wrong causal mechanism between brain areas, and on the other extreme, (2) detailed modeling with physiological details that are not needed to describe the mechanism of the recorded circuit.

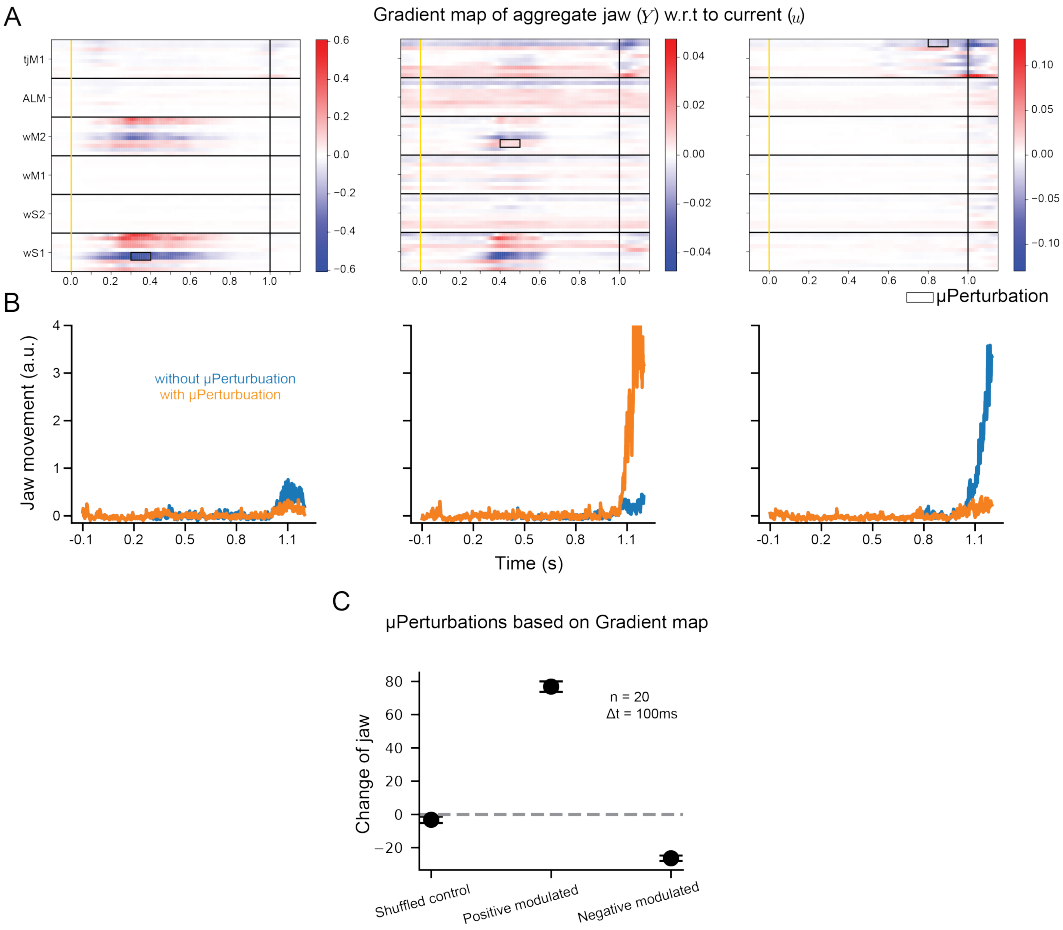

Figure 5: **Equivalence between gradients and μPerturbations A.** Mathematical gradient maps $\frac{dY}{du}$ evaluated from the model indicate how much a potential change in the input to single neurons (aligned along the vertical axis) at time t (horizontal axis) would have affected the simulated jaw movement. The open rectangular frame indicates the time window of 100ms and the set of 50 neurons selected to receive a simulated optogenetic input. Using the gradients, we selected 50 neurons that have, during that time window, a strong effect on the jaw movement in B. **B.** Actual effect of a **μPerturbations** on the jaw movement (vertical axis) in a simulated trial with the same initial conditions as in A. Left: jaw movement is further suppressed in a miss trial by stimulating excitatory neurons in wS1 at 300ms after trial onset. Middle: a miss trial is transformed into a hit trial by stimulating excitatory neurons in wM2 400ms after trial onset. Right: a hit trial is transformed into a miss trial by stimulating inhibitory neurons in tjM1 800ms after trial onset. **C.** Quantitative effect of gradient-selected μPerturbations. When the perturbations is targeted using gradient information, the perturbation effect is stronger and consistent with the gradient sign in the targeted neurons.

Our method relies on the co-development of recurrent neural network reconstructions (ReconRNN) combined with *perturbation tests* to validate data-constrained models. 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 competing models to in vivo optogenetic interventions. In future work, with higher recording density and better model design, the method is likely to capture finer temporal and spatial time scales. There is a dense knowledge of the physiological 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 rigorous *perturbation tests*, our approach offers an opportunity to test which one of these biologically informed features has a causal effect. So far, we found unambiguously that anatomically informed sign and connectivity constraints for dominant excitatory and inhbitiory cell types improve the model robustness to optogenetic perturbations.

In hindsight, it becomes apparent that biological inductive biases can be beneficial improve robustness to perturbation. This general statement is consistent with emerging results elsewhere in neuroscience: biologically inspired topological maps improves 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 potential of these models to reverse engineer the recorded system, we show that model reconstructions generate detailed sensitivity maps from passive recordings only. Moreover these maps are theoretically equivalent with functional gradients in the recorded dynamical system. We show in particular that gradients in the reconstructed model can be used to make targeted micro-stimulation in vivo. In perspective with the foundational role of gradients in machine learning theory, it 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 for 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 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 the 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 to separate the feedforward (H1) and recurrent hypotheses (H2), we construct two reference spiking circuit models RefCirc1 and RefCirc2. The two networks consits of two areas A and B, and their activity follow the hard causal inference problem from method 4.1, making hard to distinguish A1 and A2 when recording the co-activation of A and B. Moreover, to make the problem even harder, the two network 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 circuits model 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 choosen 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, this 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 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 1). 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 $\pi : \{1, \dots K\} \rightarrow \{1, \dots K\}$. 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 ReconRNNs, 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 it 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 it's 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 (see 4.7) 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 doing 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 S2. For the real dataset, we use the same value λ_3 as the one we find for the full reconstruction method of ReconRNN1.

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 ReconRNNs, 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 ReconRNN1. 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 ReconRNNs, we evaluate the effect of the perturbations directly from the neural activity. We use the distance of network dynamics φ to compare the two perturbed networks (see 4.7). 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.

4.7 Evaluate goodness-of-fit for single-trials

We compare the trial-to-trial activity of our model and the dataset using the distance of network dynamics (φ). Based on the single-trial population-average activity of the network ($\mathcal{T}'^{t}_{\text{trial},k}(\boldsymbol{z})$) and the recordings ($\mathcal{T}'^{t}_{\text{trial},k}(\boldsymbol{z}^{\mathcal{D}})$), we calculate the mean Pearson correlation coefficient ρ of the best-matched trials $(k, \pi(k))$, this matching is implemented with the Hungarian algorithm, as in the trial-matching loss. Defining $y^t_k(\boldsymbol{z}) = \mathcal{T}'^{t}_{\text{trial},k}(\boldsymbol{z}) - \langle \mathcal{T}'^{t}_{\text{trial},k}(\boldsymbol{z}^{\mathcal{D}})\rangle_t$, we can write the trial-matched Pearson correlation as:

$$\rho(z, z^{\mathcal{D}}) \quad = \quad \frac{1}{K}\sum_k \frac{\sum_t y^t_k(\boldsymbol{z}) y^t_{\pi(k)}(\boldsymbol{z}^{\mathcal{D}})}{\sqrt{\sum_t (y^t_k(\boldsymbol{z}))^2 \sum_t (y^t_{\pi(k)}(\boldsymbol{z}^{\mathcal{D}}))^2}}, \tag{10}$$

Because this measure is sensitive, when the activity is stereotypical during optogenetic activations, we use the difference between the trial-matched Pearson correlation of the recordings with itself minus the model with the recordings, $\varphi(z, z^{\mathcal{D}}) = \rho(z^{\mathcal{D}}_1, z^{\mathcal{D}}_2) - \rho(z^{\mathcal{D}}, z)$, the subscripts of $z^{\mathcal{D}}_1$ and $z^{\mathcal{D}}_2$ indicate different subsets of the dataset.

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.

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.

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.

Reimann, M. W., Bolanõs-Puchet, S., Courcol, J.-D., Santander, D. E., Arnaudon, A., Coste, B., Delalondre, F., Delemontex, T., Devresse, A., Dictus, H., et al. (2024). Modeling and simulation of neocortical micro-and mesocircuitry. part i: Anatomy. *eLife*, 13.

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.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., et al. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*, page 407007.

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.

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.

Supplemental information

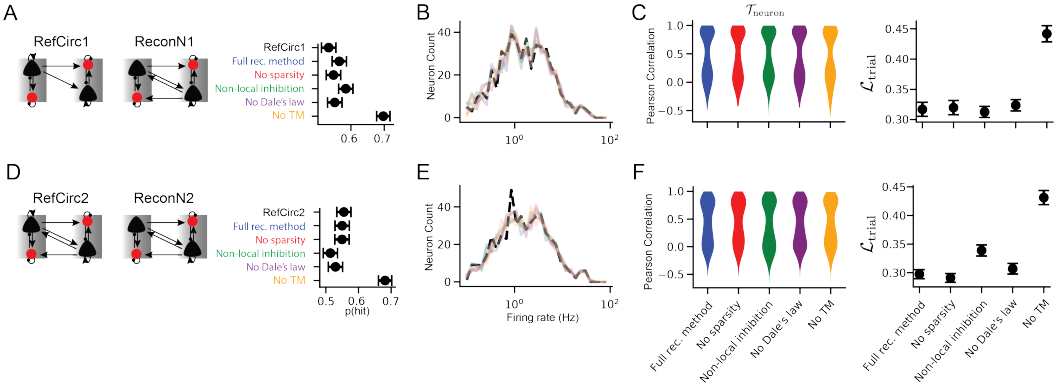

Figure S1: **Fitting Reconstructed networks to the synthetic dataset. A.** Schematic representation of the RefCirc1 and ReconRNN1. and probability of hit trials. **B.** Histogram of the firing rate distribution of the RefCirc1 and all the ReconRNN1 versions. We observe that all ReconRNN1 versions fit well with the RefCircs. **C.** Left: Violin plot of the Pearson correlation of the trial-averaged signal $\mathcal{T}_{\text{neuron}}$ for each neuron from the model and the recordings. All ReconRNN versions have a similar distribution to the RefCirc. Right: Trial-matching loss of the different ReconRNN1 variants. We observe that the model without the trial-matching loss function behaves considerably worse. The whiskers show the standard error of the mean across trials. **D-F.** Same as **A-B** for RefCirc2 and ReconRNN2.

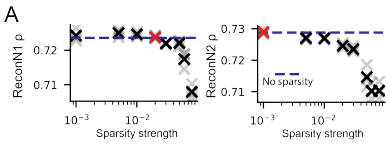

Figure S2: **Picking the sparsity level. A.** Grid search for the optimal maximum regularization strength (λ_3) without a drop in performance. As a performance measure, we used the trial-matched Pearson correlation, ρ.

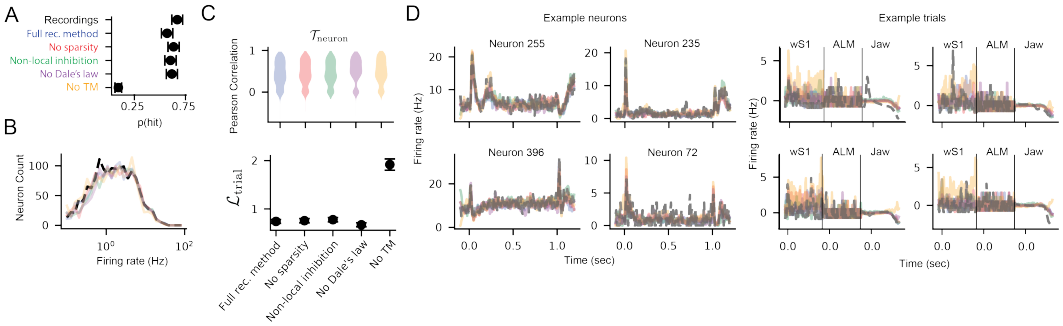

Figure S3: **Reconstruction of the real recordings. A.** Probability of hit trials of the variant reconstruction methods. **B.** Histogram of the firing rate distribution from the real recordings and all the variants. **C.** Top: Violin plot of the Pearson correlation of the trial-averaged signal $\mathcal{T}_{\text{neuron}}$ for each neuron from the model and the recordings. The variants without Dale's law and without trial matching perform better since their optimization is simpler. 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 standard error of the mean across trials. **D.** Left: The trial-averaged activity from four single neurons from the recordings and all the variant models. Bottom: The normalized population-averaged activity $\mathcal{T}_{\text{trial}}$ from four trials from the recordings and all the variant models. Here, we show only the areas wS1 and ALM because these are the only simultaneously recorded trials from session 1.

Method name	RefCirc1 vs ReconRNN1	RefCirc2 vs ReconRNN2
Full rec. method	0.46 ± 0.02	0.49 ± 0.02
No sparsity	0.44 ± 0.02	0.49 ± 0.02
Non-local inhibition	0.45 ± 0.02	0.43 ± 0.02
No Dale's law	0.44 ± 0.02	0.45 ± 0.02
No TM	0.43 ± 0.02	0.46 ± 0.02

Table 1: Trial-averaged Pearson correlation of the different reconstruction methods with the reference circuit. ± indicates the standard error of the mean across neurons. All the numbers are produced from a random model from each category.

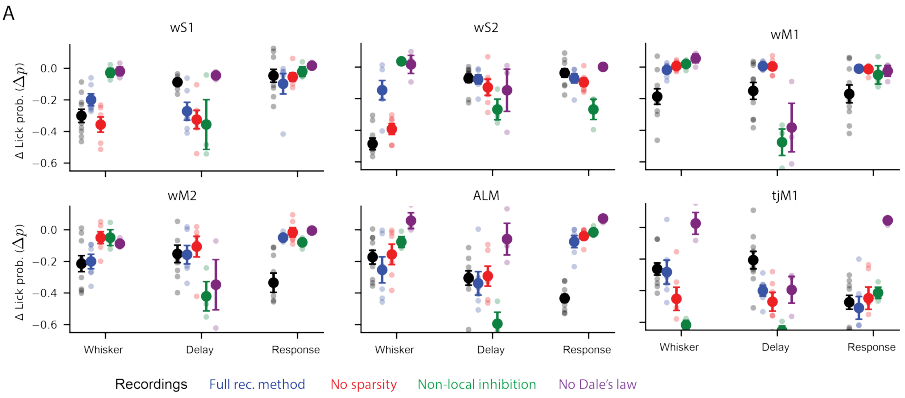

Figure S4: **A** Change of lick probability under inactivation of all areas in all the different temporal windows. We show the Δp from the data and reconstruction model variants.

Method name	Real Dataset vs Reconstructed network
Full rec. method	0.44 ± 0.01
No sparsity	0.45 ± 0.01
Non-local inhibition	0.45 ± 0.01
No Dale's law	0.52 ± 0.01
No TM	0.49 ± 0.01

Table 2: Trial-averaged Pearson correlation of the different reconstruction methods with the real recordings from (Esmaeili et al., 2021) ± indicates the standard error of the mean across neurons. We observe that the models without Dale's law and the model without trial-matching are performing better. This can be justified since optimizing these two cases is much easier.