NEUROSCIENCE

# Constructing biologically constrained RNNs via Dale's backpropagation and topologically informed pruning

**Aishwarya Balwani[1]*, Alex Q. Wang[2], Farzaneh Najafi[3], Hannah Choi[4]***



**Recurrent neural networks (RNNs) have emerged as a prominent tool for modeling cortical function. However, their conventional architecture is fundamentally lacking in physiological and anatomical fidelity, often raising questions regarding the validity of the insights gleaned from them. Our work therefore develops mathematically grounded methods that let us simultaneously incorporate Dale's law with highly sparse connectivity motifs into the RNN training pipeline such that the performance of our constrained models empirically matches that of RNNs trained without any constraints. We subsequently demonstrate the utility of our methods for inferring multi-regional interactions by training RNN models with data-driven, cell type–specific connectivity constraints to reconstruct two-photon calcium imaging data during visual behavior in mice spread across multiple cortical layers and brain areas. The interactions inferred by our models corroborate experimental findings in agreement with the theory of predictive coding, across both long and short timescales.**

## INTRODUCTION

Recent years have seen the increasing adoption of artificial neural networks (ANNs) for modeling brain function both mechanistically and algorithmically (*1*–*5*). In particular, recurrent neural networks (RNNs) are now an established tool in computational neuroscience research (*6*, *7*), being used to study neuronal computation at varying scales ranging from subsets of neurons sampled across a single brain region, two interacting regions (*8*–*10*), and even numerous populations spread across multiple interacting brain regions (*11*, *12*). By way of either reproducing desired behaviors (*13*–*15*), task-driven responses (*16*–*18*), or fitting to recorded neural data (*19*, *20*), RNNs have been shown to successfully capture latent dynamics typical of neural circuits (*21*, *22*), thus making them especially useful for modeling phenomena observed across the cortex. The degree to which ANNs can effectively approximate neural data, however, depends on two key considerations: (i) The literature suggests a direct correlation between the ability of an ANN to learn well on a task and the extent to which its behavior and learnt representations match real neural data (*23*–*25*), and (ii) more biologically realistic architectures aid in the learning of representations that better match real neuronal data (*26*–*29*). These factors make it essential that the ability of the ANNs to learn and represent a wide range of function classes be unrestricted, that their training not suffer from hindrances, and that their construction respect important anatomical principles, especially when being used as models of the brain to study neuroscientific phenomena (*30*, *31*).

Of the various discrepancies conventional RNN-based neuroscientific models have with their biological counterparts (Fig. 1, left) (*32*), two notable ones are their lack of adherence with Dale's principle (*33*), i.e., the phenomena that restrict presynaptic neuron to have exclusively either an excitatory or inhibitory effect on all its postsynaptic connections, and structured sparse connectivity among neuronal populations, a fundamental feature of brain organization observed across various species (*34*–*36*) and brain regions (*37*). Unfortunately, directly incorporating these constraints oftentimes decreases the capacity and flexibility of the network to fit the training data, which leads to a drop in learning performance (*38*, *39*). While there has been active research toward addressing these issues in both the machine learning (ML) (*40*–*43*) and computational neuroscience communities (*38*, *44*–*49*), these efforts have mostly been made to include these constraints into the network structure individually, rather than in conjunction as one would see in biological brains (*50*). Subsequently, there remains a need for ways to construct sparse, sign-constrained deep neural networks that can also achieve performance levels comparable to conventional ANNs.

Our work therefore introduces methods that allow us to easily incorporate both neuronal sign constraints and sparse connectivity motifs into the conventional backpropagation-based RNN training pipeline (Fig. 1A). Specifically, we first train a dense network that respects a predetermined set of sign constraints via a modified version of standard backpropagation (*51*), which we call Dale's backpropagation (Fig. 1B), after which we prune away weights using a probabilistic pruning rule we call top-prob pruning (Fig. 1C), to achieve a target connectivity pattern. Finally, we retrain the sparse sub-network retained post-pruning once again with Dale's backprop. Both of our methods are mathematically grounded and follow rigorous principles. With Dale's backprop, we provide theoretical guarantees on the linear convergence of the algorithm under specific conditions, ensuring that the training respects anatomical constraints while achieving optimal learning performance. Our pruning rule is motivated by topological principles, particularly the preservation of high-magnitude weights that contribute to the network's zeroth-order connectivity structure, thus enhancing the functional and anatomical plausibility of the model.

Besides being easily implementable and scalable using standard ML packages, our approach also aligns with the biological processes of synaptic development and refinement. Given that synaptic connections initially form abundantly with many connections later being pruned based on activity and functional relevance, by first learning a dense set of weights with Dale's backprop the network can capture a rich set of connections that adhere to Dale's law, reflecting

[1]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA. [2]Computational Science and Engineering Program, Georgia Institute of Technology, Atlanta, GA, USA. [3]School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, USA. [4]School of Mathematics, Georgia Institute of Technology, Atlanta, GA, USA.
*Corresponding author. Email: abalwani6@gatech.edu (A.B.); hannahch@gatech.edu (H.C.)

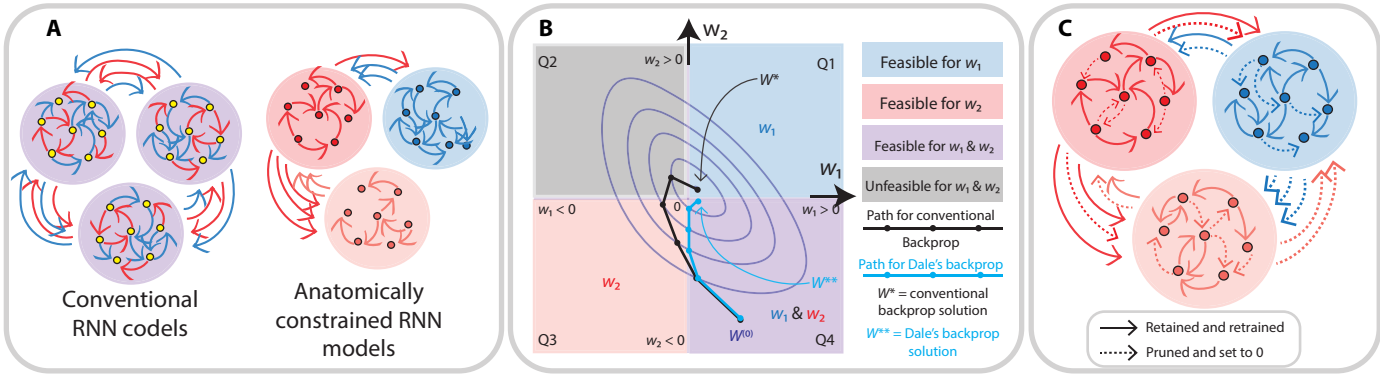


**Fig. 1. Schematic for constructing biologically constrained RNN models.** (**A**) Illustration of conventional versus biologically constrained RNN models. Conventional RNNs consist of general purpose neurons that project a mix of excitatory and inhibitory signals, with no specific connectivity structure within or across populations. Biologically constrained RNNs restrict populations of neurons to be either strictly excitatory (red) or inhibitory (blue), with anatomically informed connectivity motifs both within and across populations. (**B**) Optimization in parameter space when training with conventional backpropagation (black) versus Dale's backprop (blue). Purple contours represent level sets for the positions the algorithms take in parameter space at different time steps. (**C**) Enforcing anatomically consistent connectivity motifs. Dashed lines represent connections that are set to 0 during the pruning process. Solid lines represent connections that are retained post-pruning.

excitatory and inhibitory roles at a fundamental level. Subsequent application of top-prob pruning mirrors the refinement phase, where weaker, less functionally critical synapses are eliminated, retaining only the most effective pathways. This pruning rule not only emphasizes synaptic efficacy (*52*) by preserving stronger synapses but also adheres to principles of synaptic scaling (*53*) in the retraining phase by maintaining a balanced level of activity within the network—as during the retraining phase, the synaptic strengths within the network are dynamically adjusted, effectively rescaling the remaining connections to maintain overall network activity and prevent neuron underutilization. This mirrors the biological process of synaptic scaling, ensuring that the network retains its capacity to learn and generalize despite the reduced number of connections. Overall, this process of initial dense learning followed by selective refinement embodies how biological systems evolve and ensures computational efficiency by optimizing for both anatomical and functional plausibility (*54*, *55*).

We demonstrate the suitability of our methods for studying neuroscientific data by applying them on RNNs learning a two-photon calcium imaging dataset, exploring multi-regional interactions that underlie visual behavior in mice when performing a change detection task (*56*). We find that our models successfully recapitulate both long- and short-timescale interactions among neuronal populations, capturing transient dynamics as well as sustained signals that are critical for complex perceptual processing. Moreover, our model outputs align with the predictive coding hypothesis (*57*), as they reflect anticipatory and feedback-driven patterns observed experimentally, suggesting that our approach is well suited to modeling the layered processing of sensory information in the brain.

Together, our results on synthetic and real-world datasets indicate that our methods offer a robust framework for fitting and modeling neural dynamics in a biologically faithful manner. By capturing both anatomically realistic connectivity patterns and functional interactions, our approach provides a set of powerful tools for understanding the complex, hierarchical processing of information across different cell types, populations, and brain areas. These tools subsequently enable models to better reflect anatomical structures, thereby imparting greater confidence in their findings and enhancing the alignment between RNNs and real neural circuitry.

## RESULTS

### Dale's backpropagation enables training of sign-constrained RNNs

Here, we introduce our sign-constrained learning rule, Dale's backpropagation, and validate its performance. We provide intuition for the algorithm, present the statements of the theoretical analyses performed (i.e., convergence guarantees and error bounds), and provide empirical results demonstrating its utility on a set of neuroscience-inspired and ML tasks.

### Dale's backpropagation: Sign-constrained weight update

Dale's backpropagation enforces Dale's principle by integrating sign constraints into the conventional backpropagation process. Specifically, it employs a projection step (similar to that of projected gradient descent) on the learnt parameters at every iteration to ensure that the weights remain non-negative for excitatory neurons and nonpositive for inhibitory neurons, thus adhering to biological constraints (Fig. 1B). Combined with the thresholding of the hidden states of the RNN so that they are always non-negative, this projection step ensures that the action of every neuron respects the sign constraints of our choosing. The projection step itself can be easily implemented using masking matrices as follows

$$\mathbf{W}_{\mathrm{D}}^{(i)} = \max\left(0, \mathbf{W}_{[N^+]}^{(i)}\right) \oplus \min\left(0, \mathbf{W}_{[N^-]}^{(i)}\right)$$

where $\mathbf{W}_{\mathrm{D}}^{(i)}$ represents the weight matrix that satisfies Dale's law at iteration $i$ and weight subsets corresponding to excitatory and inhibitory neurons are denoted by $[N^+]$ and $[N^-]$, respectively. The technical details and added intuition for the derivation of this rule are provided in Materials and Methods.

### Theoretical results

In this part of the paper, we intuitively discuss our key mathematical analyses for Dale's backprop with ReLU nonlinearities when it utilizes gradient descent as its optimizer. First, we derive the rate of convergence of the algorithm under the assumption of restricted optima, i.e., when we can assume that the optimal set of parameters also has the same sign pattern as that imposed. Second, we quantify the differences between Dale's backprop and standard backpropagation,

in terms of both the weights learnt and the final solutions obtained. Together, these results establish a solid theoretical foundation for Dale's backprop, demonstrating its ability to learn effectively and efficiently, thus validating its use in modeling neural data. For completeness, we provide the theorem statements here in the main text, while deferring the lemmas used to prove the theorems, as well as the full proofs, to the Supplementary Materials (section S3.1)

We start by examining the behavior of Dale's backpropagation algorithm under the assumption that the optimal set of parameters for a task shares the same sign pattern as the one imposed—a condition we refer to as the restricted optima assumption—and show that despite having to learn with constraints, under this assumption, Dale's backprop converges linearly to the optimal solution (Theorem 1). Biologically, this assumption mirrors the idea that the arrangement of excitatory and inhibitory neurons in the network is optimized for such tasks.

Proving this theorem relies on the geometric observation that the restricted optima assumption ensures that the globally optimal set of weights ($\mathbf{W}^*$) lies within the same orthant as our point of initialization $\left(\mathbf{W}^{(0)}\right)$. We subsequently prove optimal sign pattern preservation (Lemma 5), which guarantees that every backpropagation iteration never leaves this orthant, implying that the signs of the weights remain constant throughout the optimization process. As a result, Dale's backpropagation behaves identically to unconstrained gradient descent within this orthant, making the projection step redundant since the optimization path does not approach the boundaries of the orthant. Consequently, the algorithm can take the most direct path to the optimum without any detours induced by constraint enforcement, allowing it to achieve a linear convergence rate under the Polyak-Łojasiewicz condition.

**Theorem 1 (convergence of Dale's backpropagation).** Let $\ell$ be a loss function satisfying the μ–Polyak-Łojasiewicz condition, with gradients that are $L$-Lipschitz such that $L \geq \mu > 0$. Consider the sequence of weights $\left\{\mathbf{W}_{\mathrm{D}}^{(i)}\right\}$ generated according to the Dale's backpropagation update, with a step size of $\frac{1}{L}$. Given an optimal loss $\ell^* = \ell(\mathbf{W}^*) = \operatorname{argmin} \ell\left(\mathbf{W}_{\mathrm{D}}\right)$ where $\mathbf{W}^*$ has the same sign pattern as all $\mathbf{W}_{\mathrm{D}}^{(i)}$ and a specific error $\varepsilon > 0$, it holds for iteration $i$ that

$$\ell\left(\mathbf{W}_{\mathrm{D}}^{(i)}\right) - \ell^* \leq \varepsilon \text{ when } i \geq \frac{\log\left(\frac{\ell\left(\mathbf{W}_{\mathrm{D}}^{(0)}\right)-\ell^*}{\varepsilon}\right)}{\log\left(\frac{L}{L-\mu}\right)}$$

Notably, our analysis reveals that under the assumption of the restricted optima condition, Dale's backpropagation achieves a linear convergence rate that matches the performance of unconstrained backpropagation (*58*). This is of consequence, as it demonstrates that under the right conditions, imposing biological constraints through Dale's principle does not necessarily come at the cost of convergence speed. Furthermore, it also suggests that the brain's neural circuitry despite being constrained by Dale's law might also be functionally organized to facilitate efficient learning and task performance. However, it is important to note that these guarantees rely not only on the restricted optima assumption but also on the gradients satisfying Lipschitzness and the Polyak-Łojasiewicz condition, which from a mathematically rigorous perspective may not always hold in practice.

Additionally, Dale's backprop also lends itself well to analyzing its behavior with respect to standard backpropagation when we do not make the restricted optima assumption. Specifically in the case of a single-layer RNN (without biases), we can characterize the distance between the weights found using standard backprop and Dale's backprop (Lemma 7), and therefore subsequently the distances between outputs found using the two weight update schema, allowing us to bound the difference between the final error of the solution found using Dale's backprop in terms of that found using standard backprop (Theorem 2).

The lemma on the distance between learnt weights quantifies how Dale's backpropagation diverges from standard backpropagation over time due to the sign constraints, showing that this divergence grows but remains bounded, influenced by factors like the learning rate, the loss landscape's smoothness, and gradient magnitudes. Building on this, the theorem on error between solutions relates the performance of Dale's backpropagation to standard backpropagation, indicating that the error of Dale's method is bounded by both the divergence of the weights (bounded by the lemma) and the sensitivity of the network's output to weight changes, alongside the error of the standard method. Together, these results provide theoretical assurances that, while biological constraints affect learning dynamics, they do not cause uncontrolled error growth, supporting the use of Dale's principle in neural network training. Formally, we express this theorem as follows:

**Theorem 2 (differences in errors between solutions).** Let $f(\mathbf{W})$ be the function represented by a single-layer RNN unrolled over $T$ time steps, with weights $\mathbf{W}$. Let $\mathbf{W}_{\mathrm{D}}$ be the weights learnt using Dale's backpropagation, and $\mathbf{W}$ be the weights learnt using standard backpropagation. Assume the nonlinearity $\phi$ is either tanh or ReLU. Then, the error of the solution found using Dale's backpropagation with respect to the ground truth $y$ is bounded by

$$\left\|f\left(\mathbf{W}_{\mathrm{D}}\right) - y\right\|_2^2 \leq 2 \cdot \left[\delta^2 \sum_{t=1}^{T} \left(L_{f_t}\right)^2 + \sum_{t=1}^{T} \left(\varepsilon_t^*\right)^2\right]$$

where $f\left(\mathbf{W}_{\mathrm{D}}\right)$ is the output after $K$ training iterations and $\delta = \frac{G}{L}\left[(1+\eta L)^K - 1\right]$, $L_{f_t} = \max\left(L_{f_t(\mathbf{W})}, L_{f_t(\mathbf{W}_{\mathrm{D}})}\right)$ is the maximum of the Lipschitz constants of the two RNNs at time step $t$, and $\varepsilon_t^* = \|f_t(\mathbf{W}) - y_t\|$ is the error of the solution found using conventional backpropagation at time step $t$.

Note that $\|\cdot\|_2$ always corresponds to the operator norm $\|\cdot\|_{\mathrm{op}}$ induced by the 2-norm, which is the Euclidean norm if $\mathbf{W}$ and $\mathbf{W}_{\mathrm{D}}$ are vectorized, and $\sigma_{\max}(\cdot)$, i.e., the largest singular value of the matrix if $\mathbf{W}$ and $\mathbf{W}_{\mathrm{D}}$ are considered in their matrix forms. As before, the statement of the lemma as well as full proofs for both the lemma and theorem are provided in the Supplementary Materials (section S3.2).

## Empirical results on standard tasks

We evaluate the performance of Dale's backpropagation across three tasks of interest (Fig. 2A). The first is a 1-bit flip-flop task (Fig. 2A, top row, left), in which the network is required to maintain and toggle between different states in response to a series of binary inputs. Specifically, the network output is meant to start at zero, following which it takes the value ±1 to match the input signal whenever presented. It is then expected to switch signs if presented with a signal of the opposing sign, or else, maintain the same output as before. Next, we tested a wave reconstruction task (Fig. 2A, bottom row) where both the excitatory and inhibitory neurons are presented with separate sinusoidal waveforms. The network is tasked with accurately

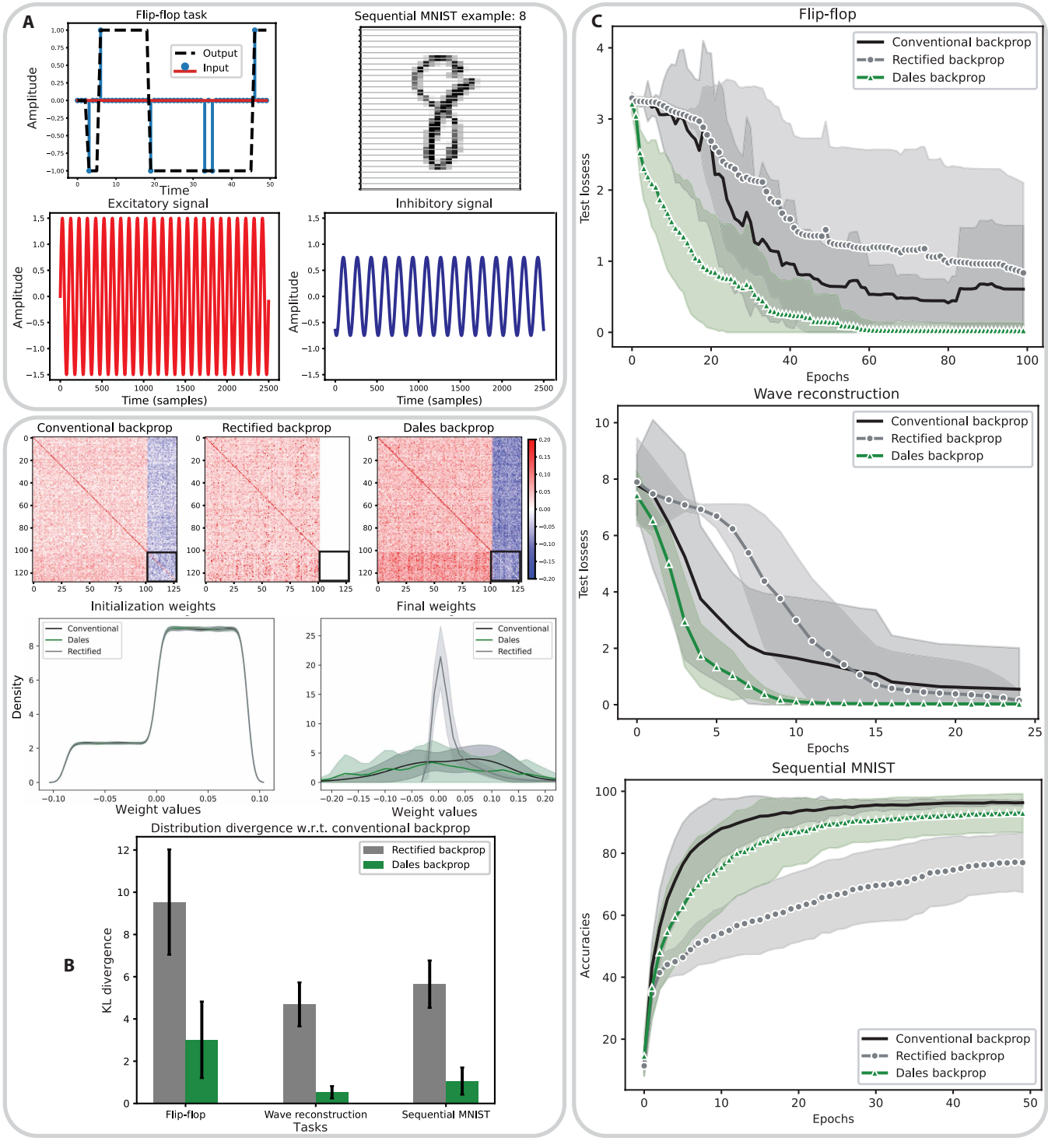


**Fig. 2. Training with Dale's backprop.** (**A**) Task examples: 1-bit flip-flop, sequential MNIST, and wave reconstruction. (**B**) Distribution of weights: Examples of weight matrices post-training (top row), weight histograms at initialization and after training (middle row), and relative divergence in weight distributions with rectified and Dale's backpropagation versus conventional backpropagation (bottom row). (**C**) Test performance of models across different tasks when trained with conventional backpropagation (black), rectified backpropagation (gray), and Dale's backpropagation (green). All statistics computed over 20 independent runs.

reconstructing both signals simultaneously, reflecting the roles of excitation and inhibition in modulating distinct aspects of signal processing in neural circuits. We finally also test our methods on the sequential MNIST task (Fig. 2A, top row, right), which is a variation of the classical digit classification task in that instead of receiving the entire image as the input, the network instead sequentially receives the rows of the image.

For each of the tasks, we train RNNs (over five runs, unless stated otherwise) with 128 hidden neurons, of which ~80% (102) are excitatory and ~20% (26) are inhibitory. In addition to conventional backpropagation, we also include in our experiments "rectified backprop" (*45*, *50*), an alternative sign-constrained method for enforcing Dale's principle. Rectified backprop works by simply zeroing out all negative weights (e.g., using the ReLU activation function) after each standard backpropagation update. Unlike our approach, which uses a principled projection operation that minimizes distance to the original update in parameter space, rectified backprop discards gradient information for inherently negative weights and instead always uses the gradient from only those that are positive. This approach has limitations—particularly its incompatibility with activation functions that produce negative outputs (such as tanh) and its tendency to create "dead neurons" (i.e., neurons with projections nearing zero such that their effect on the overall computations of the network is negligible) as demonstrated in our results (Fig. 2B, top row, middle). We include this comparison because, like our method, rectified backprop can be combined with sparsity constraints, whereas other approaches for implementing Dale's principle (*38*, *59*) cannot be easily adapted to respect structured sparse connectivity motifs, making them unsuitable for our objective of simultaneously enforcing both Dale's law and anatomically informed sparsity patterns.

Our experiments justify our proposed weight update and subsequent theoretical analyses, in terms of how the weights evolve, as well as learning performance. We start by analyzing the distribution of weights before and after training for the three learning rules (Fig. 2B) averaged across all tasks. While we initialize all three methods identically (Fig. 2B, middle row, left), we notice that their distributions post-training are visibly different (Fig. 2B, middle row, right). Specifically, the weights learnt using conventional backprop (black) show a small peak around zero and a large deviation, similar to those learnt using Dale's backprop (green), but the weights learnt using rectified backprop show a sharper peak around zero (gray). This trend is also visible in the weight matrices themselves (Fig. 2B, top row), where the negative weights learnt using rectified backprop are practically zero (Fig. 2B, top row, middle). On first glance, the weight matrices for conventional and Dale's backprop seem almost the same, but closer inspection (black boxes, bottom right of the weight matrices) reveals that some of the weights that should have been negative have flipped signs with conventional backprop (Fig. 2B, top row, left) but there are no such discrepancies with Dale's backprop (Fig. 2B, top row, right). Finally, we empirically quantify the differences in learnt weights for the two sign-preserving methods by measuring the Kullback-Liebler (KL) divergence among their weight distributions post-training (Fig. 2B, bottom row) with respect to conventional backpropagation. We observe that across all three tasks the divergence shown by rectified backprop (gray) from the weights learnt by conventional backprop is significantly higher than that shown by Dale's backprop (green), with $P$ values of $3.418 \times 10^{-3}$, $3.223 \times 10^{-13}$, and $2.316 \times 10^{-8}$, respectively, for the flip-flip, wave reconstruction, and sequential MNIST tasks.

Finally, the learning performance (Fig. 2C) of Dale's backprop (green) matches that of conventional backprop (black) for all three tasks, at times even learning faster. We conjecture that this is a consequence of the regularization introduced by adhering to the sign constraints—by restricting the optimization to the orthant where the signs are preserved, the search space is effectively reduced. This focused parameter space allows for more efficient learning dynamics, as the optimizer concentrates on adjusting weight magnitudes without expending effort on sign changes that violate the constraints. The fixed signs lead to more stable and directed weight updates, resulting in a smoother optimization landscape. Consequently, Dale's backprop can converge more rapidly in the early stages of training while ultimately achieving similar final performance as standard backpropagation. However, the learning performance of rectified backprop (gray) is both slower and not as competitive as the other two methods, especially on the more complicated sequential MNIST task. To that end, we note that this might be a consequence of the fact that rectified backprop does not allow for activation functions that have any negative outputs (e.g., tanh). This restriction likely leads to exploding gradients and dead neurons—the latter which might also be inferred visually from its weights post-training. However, Dale's backpropagation does not suffer from such limitations and can use nonlinearities that have negative outputs as long as it is centered around 0, i.e., it keeps positive and negative activations, positive and negative, respectively.

## Pruning can enforce sparse connectivity across neuronal populations

Having established a method that lets us train sign-constrained RNNs leveraging the machinery of autograd-based backpropagation (*60*, *61*), in this section, we now describe topologically informed probabilistic (top-prob) pruning as a way of sparsifying dense neural networks to reflect a target connectivity pattern among neuronal populations (Fig. 1C). We state the pruning rule formally, and with subsequent empirical analyses demonstrate the applicability of our method in conjunction with Dale's backpropagation, wherein it outperforms the random pruning baseline on different tasks.

## Topologically informed probabilistic pruning rule

Our top-prob pruning rule allows us to take a densely connected network and sparsify it such that the retained connections mimic a sparse target connectivity. We do so probabilistically, where rather than using deterministic cutoffs, our pruning rule assigns each connection (i.e., learnt weight) a probability of being retained based on its strength—so stronger connections have higher chances of survival. This approach creates a natural balance where important pathways tend to remain intact, while weaker, less influential connections may be removed. Additionally, the probabilistic nature of this mechanism allows for maintaining diverse connectivity patterns that might be eliminated by strict threshold-based approaches while also allowing us to prune additional weights either single-shot or iteratively, gradually sparsifying the dense network through cycles of pruning and retraining and fine-tuning.

Consider a weight matrix $\mathbf{W} \in \mathbb{R}^{N \times N}$ composed of the synaptic weights $w_{ji} \forall i, j \in \{1,2,3, \ldots, N\}$, connecting neuron $i \rightarrow j$. The sparsified matrix $\mathbf{W}^{\text{sparse}} \in \mathbb{R}^{N \times N}$ is obtained using the pruning rule

$$w_{ji}^{\text{sparse}} = \begin{cases} w_{ji} & \text{with probability } \kappa\,|w_{ji}| \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $\kappa \in \mathbb{R}^+$ is a non-negative scalar that controls the sparsity of the resulting matrix and is defined as

$$\kappa = \frac{(1-s)N^2}{\|\mathbf{W}\|_{L^1}^2} \tag{2}$$

$s \in [0, 1]$ is the target sparsity of $\mathbf{W}^{\text{sparse}}$ and $\|\mathbf{W}\|_{L^1}^2 = \sum_{i=1}^{N} \sum_{j=1}^{N} |w_{ji}|$.

While it is evident that the top-prob pruning rule operates by probabilistically retaining weights of higher magnitude while eliminating weaker ones, we emphasize that this approach mirrors fundamental aspects of synaptic plasticity in biological neural networks. The rule's local nature—where pruning decisions depend solely on individual synaptic weights—aligns with biological constraints, as real neurons modify their connections based only on local synaptic properties rather than on global network states. Furthermore, when coupled with Dale's backpropagation, the top-prob pruning mechanism has a propensity for preserving exactly those weights that align with Hebbian learning principles (section S2), thereby ensuring the maintenance of biologically meaningful effective connectivity while simultaneously achieving network sparsification.

The justification for the rule from an ML/mathematical perspective is provided in Materials and Methods. Throughout the remainder of this work, we use top-prob pruning in the one-shot sense (*41*) wherein we prune to a target sparsity level and connectivity pattern in a single step, followed by a single retraining phase to help restore the model's performance.

## Empirical results on standard tasks

We study the behavior and performance of top-prob pruning by first examining how it affects the weight distribution and structural integrity of the original dense network, followed by its ability to maintain functional capacity. Our results suggest that top-prob pruning does indeed preserve key network properties, leading to highly sparse yet robust models that do not require extensive retraining to regain performance.

As a preliminary check, we observe the distribution of nonzero weights (Fig. 3A) for dense weight matrix (black) versus that of a matrix that has been pruned to 90% sparsity using the top-prob pruning rule (green) and random pruning (gray). As expected, we see that the dense matrix has weights that are almost uniformly distributed since the weights were sampled from the distribution $\mathrm{U}\left(\left[\frac{-1}{\sqrt{N}}, \frac{1}{\sqrt{N}}\right]\right)$ and the randomly pruned network stays faithful to this distribution. The weights retained using top-prob pruning, however, are heavily skewed toward being higher in magnitude, demonstrating that it successfully prioritizes stronger connections while eliminating weaker ones. We subsequently quantify the notion of

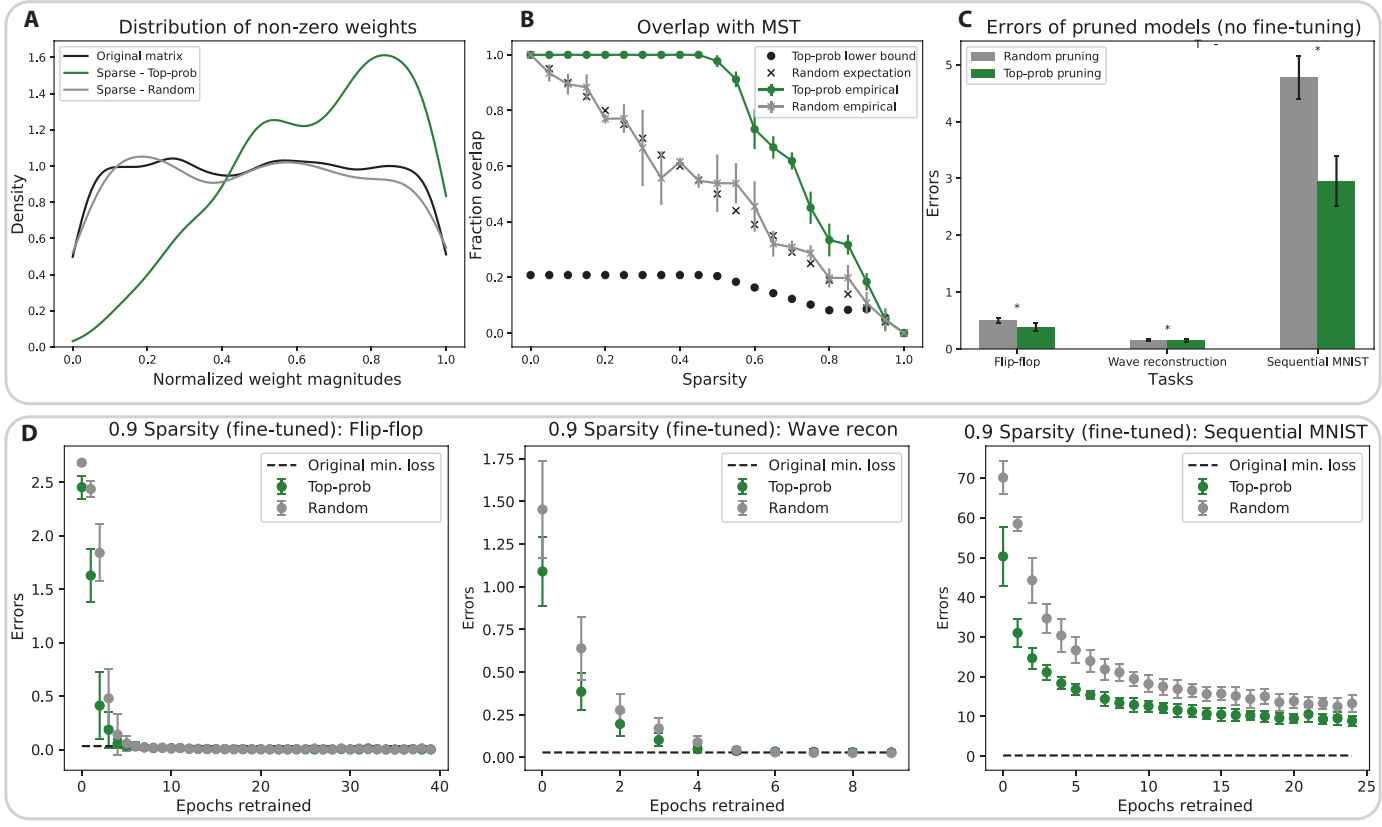


**Fig. 3. Topologically informed probabilistic pruning.** (**A**) Distribution of nonzero weights for dense (black) and sparse matrices pruned randomly (gray) and with top-prob pruning (green). (**B**) Fraction overlap that retained weights have with the MST of the dense matrix. The "Top-prob lower bound" represents a loose theoretical minimum overlap we would expect between our top-prob pruned network and the MST, calculated based on the probability that high-magnitude weights (which our method preferentially retains) are also part of the MST structure. More details are available in section S5.1. (**C**) Errors of pruned models, without any retraining using random (gray) and top-prob pruning (green). (**D**) Performance of sparsified and fine-tuned models across different tasks, when pruned randomly (gray) versus top-prob pruning (green). All statistics computed over five independent runs. *P ≤ 0.05.

"structural integrity preservation" by plotting the fraction overlap of retained weights with the maximum spanning tree (MST) of the dense matrix as sparsity increases (Fig. 3B) for both pruning methods. The MST is a subset of connections that connects all nodes in a graph while maximizing the total connection weight. In our case, this weight is typically defined as the magnitude of synaptic strength, ensuring that the strongest connections are prioritized in the spanning tree. Here, MST represents the essential connectivity backbone of the original dense network in the topological sense (*62*), and a higher overlap with the dense network's MST indicates that the pruning method is better at preserving these critical pathways that maintain the network's fundamental structure and information flow capabilities. Again, we observe that empirically top-prob pruning shows a much higher overlap with the MST (green dots) compared to random sparsification in practice (gray crosses) and theoretically (black crosses) (section S5.2). Additionally, it shows lesser variations in the amount of overlap as well.

Moving on to the pruned network's ability to retain information and functional capacity, we find that networks pruned with our top-prob method consistently outperform randomly pruned networks. Specifically, the prediction errors immediately after pruning (before any retraining; Fig. 3C) are always lower for top-prob pruned models compared to randomly pruned models at 90% sparsity. Moreover, the difference in errors becomes more apparent as the task complexity increases ($P = 0.022$ for the flip-flop and wave reconstruction tasks, while $P = 0.012$ for sequential MNIST). Fine-tuning with Dale's backpropagation for ~50% the number of epochs as the original training while retaining the sparse structure identified via pruning follows a similar trend (Fig. 3D), with networks that that were pruned randomly (gray) showing higher errors at the epoch of retraining than those pruned with top-prob pruning (green). While both methods lead to models that seem to eventually approach the original model's performance (dashed line) after fine-tuning, top-prob pruning consistently starts from a better initial error, converges faster to optimal performance, and shows more stable learning. We therefore conclude that our pruning rule effectively identifies and retains functionally important weights. The preserved connectivity aligns well with the network's core topology (MST), which leads to efficient and robust performance of the sparsely structured RNN, in conjunction with Dale's backpropagation.

## Application to infer effective connectivity of cortical populations during visual behavior in mice

Having established the efficacy of our methods in successfully constructing and training highly sparse RNNs that respect Dale's law, we apply them to study interactions among neural populations during visual behavior in mice under the predictive coding hypothesis (*57*, *63*). Specifically, we model data from the Allen Institute Visual Behavior dataset (*64*, *65*), which comprises two-photon (2p) calcium imaging recordings from mice performing a change detection task, when presented with expected and unexpected stimuli. This experimental paradigm allows us to study how neural populations interact in the mouse visual cortex in a predictive context. Going a step further from the work of (*20*), which used data-constrained but otherwise conventional Elman-style RNNs to infer macroscopic brain-wide interactions, we train anatomically constrained RNNs on neural activity data. Our anatomically constrained models learn to not only imitate neural dynamics but also do so with biologically realistic neural circuits, revealing how effective connectivity adapts within the network at the level of brain areas, layers, and cell types. This approach provides concrete insights into circuit-level mechanisms that are compatible with predictive coding principles.

Our results align well with previous observations made in the experimental literature studying the data and are consistent with predictions of the predictive coding hypothesis. By examining how the inferred connectivity patterns change across different experimental conditions, our approach both validates existing experimental findings and generates previously unidentified hypotheses. This allows us to identify specific feedforward and feedback pathways that engage differently based on prediction errors, revealing fundamental organizing principles of cortical circuits during predictive processing.

## Dataset and model overview

Our modeling application uses the Allen Institute's Visual Behavior dataset (*64*, *65*), which records neural activity from the visual cortex of mice performing a change detection task. The data are collected from areas V1 (primary visual cortex) and LM (lateromedial area) across cortical layers identified by recording depths, and from diverse cell types including excitatory (Pyr; pyramidal) and two types of inhibitory neurons, namely, somatostatin (Sst)– and vasoactive intestinal peptide (Vip)–expressing neurons. Subsequently, our model, the CelltypeRNN, is a biologically constrained RNN consisting of the two hierarchically related regions V1 and LM (Fig. 4A), each with the simplified structure of a cortical column and three cell types (Pyramidal, Vip, and Sst).

The mice are shown a series of familiar (Fig. 4B, top row) and novel images (Fig. 4B, bottom row) and rewarded for correctly identifying image changes. Neural responses are collected under three conditions: no change (Fig. 4C, second row), change (Fig. 4C, third row), and omission (when the image is replaced by a blank screen) (Fig. 4C, first row). This experimental design provides an ideal testbed for examining predictive coding principles, as the different stimulus conditions (familiar/novel × change/no change/omission) generate distinct types of expectation violations that should engage different aspects of hierarchical prediction and error-correction mechanisms, which are apparent in the recorded neural responses. These responses are modeled (Fig. 4C, bottom) across two temporal windows: full-set (covering two image presentations) (Fig. 4C, third row) and half-set (starting after the second image) (Fig. 4C, first row).

The CelltypeRNN model is trained using Dale's backpropagation to enforce biologically realistic weights across the various cell types. Following the first training step, top-prob pruning is applied to remove weaker connections while retaining stronger ones (Fig. 4D, left to right), thus sparsifying the model to have the target connectivities across neuronal populations specified by (*66*). Finally, the highly sparse model is fine-tuned by additional training, again using Dale's backpropagation. In both training phases, the model predicts neural activity for each population at the next time step. We train separate models for each stimulus condition ([familiar/novel] × [change/no change/omission] × [full/half set]), giving us a total of 12 different CelltypeRNN models. All models are trained over five random seeds, and results shown in the subsequent figures are averages over all runs unless stated otherwise.

Additional details for data curation, preprocessing, model architecture, and training are provided in Materials and Methods. Code to train and saved reconstruction results for various conditions are publicly available on the project GitHub at: https://github.com/HChoiLab/biologicalRNNs/tree/main/celltype-recon.

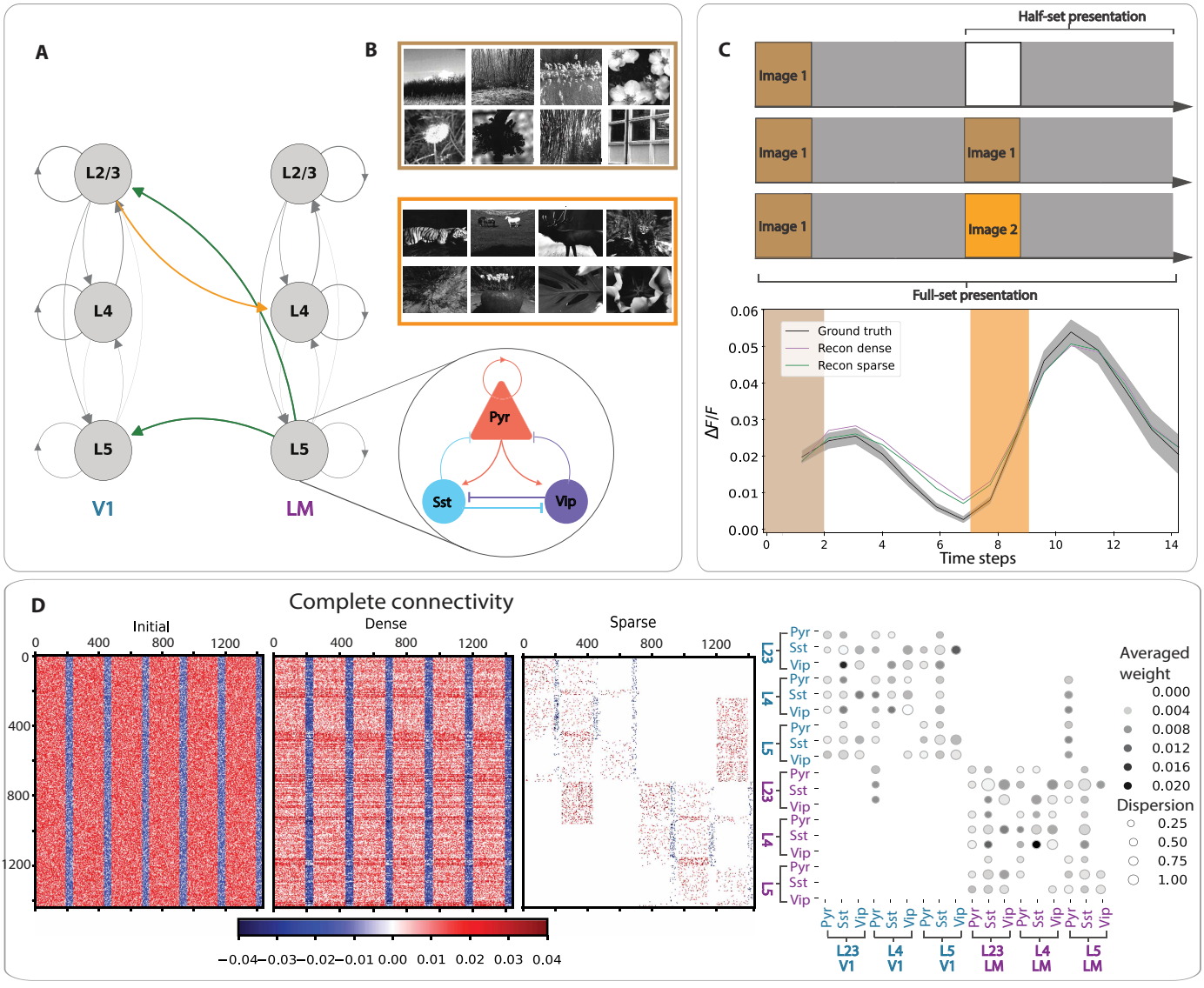


**Fig. 4. Dataset, network structure, and task schematics.** (**A**) General architecture of the CelltypeRNN. (**B**) Familiar and novel image sets used for training mice on the visual change detection task. Reproduced from the Allen Institute Visual Behavior-2p dataset (open source) (*64*). (**C**) Top: Examples of different stimuli conditions in the visual change detection task and depiction of full- and half-set presentation timescales. Bottom: An example of mean target activity (black curve), dense RNN output (purple curve), and sparse RNN output (green curve) from V1 L2/3 Pyr population in the novel change condition. The mean target activity was averaged over a randomly sampled subset of 100 neurons, selected from 9015 neurons, while the reconstructions were averaged over five different trained networks each. (**D**) Examples of inferred effective connectivity. Left: Complete neuron-to-neuron connectivity at initialization, after training with Dale's backprop, and after sparsification (and retraining) with top-prob pruning. Neurons are ordered by area (V1 followed by LM) within which they are ordered by layer (L4, L2/3, L5), and type (Pyr, Sst, Vip). Right: Example of the sparse connectivity matrix where activity is averaged by cell type in every layer (bigger circles imply higher dispersion, and darker colors imply stronger connections; dispersion is computed as the fraction of standard deviations to the mean activity in the population).

## Biologically constrained RNN models infer population interactions consistent with predictive coding

Using the anatomically constrained CelltypeRNN architecture, we examine how distinct cell types across the layers and hierarchy in the visual cortex communicate both expected and unexpected information by comparing inferred connectivity patterns across different experimental conditions and timescales.

Our approach trains the RNN to perform one-step-ahead prediction of neural activity for each of the 18 cell type–specific populations (across different cortical layers in V1 and LM). This is formalized in our loss function (*6*), which minimizes the errors between the network's outputs and the actual neural activities from the data (*64*) in the next time step. As mentioned before, we train completely separate RNN models for each of our 12 experimental conditions (familiar/novel × change/no change/omission × full-set/half-set presentation), rather than adapting a single network. This allows us to compare how the inferred connectivity differs across contexts.

By fitting neuronal responses of interacting populations through one-step-ahead predictive reconstruction, we capture the dynamic temporal dependencies inherent in neural activity. The RNN's resulting

connectivity matrix, constrained by both Dale's principle and cell type–specific connection probabilities (table S1), serves as a functional proxy for interactions among populations, reflecting how signals propagate within the cortical network. After initial training with Dale's backpropagation, we apply top-prob pruning to enforce the biologically motivated sparse connectivity patterns (as described in table S1), followed by fine-tuning to restore performance.

The rationale for this approach is that the learned weights in our model represent the functional influence that each population exerts on others. By analyzing how the RNN adjusts its connectivity across varying predictive contexts and timescales, we gain insights into circuit-level implementations of predictive coding, particularly in prediction error communication and modulation of feedforward and feedback pathways. This allows us to study effective connectivity across the circuit at both longer and shorter timescales by utilizing either sustained dynamics across entire stimulus sequences or immediate neural responses to prediction confirmations or violations.

Hierarchical predictive coding theory proposes that higher cortical areas continuously generate predictions about expected sensory inputs, which are compared against actual signals at lower processing levels. When mismatches occur—such as unexpected visual changes or stimulus omissions—prediction errors are computed and communicated through specific neural pathways: Feedforward connections carry error signals upward to update higher-level models, while feedback connections transmit revised predictions downward to modulate lower-level processing. The experimental paradigm in our dataset naturally generates such prediction violations through image changes (temporal expectation violations), omissions (stimulus expectation violations), and novel images (statistical regularity violations), allowing us to examine how these theoretical principles manifest in cortical circuit dynamics and interactions among cell populations. Our results can be broken down into three key comparisons.

***Familiar no change versus familiar change (full-set presentation)***

In the full-set presentation of familiar images, we observe substantial differences in the inter-areal feedforward and feedback connections (Fig. 5A). When the activities are averaged across layers, there is a stark increase in the projection V1 L2/3 → LM L4 when there is a change in the image compared to when there is not, suggesting that the expectation violation causes enhanced forward communication from V1 to LM (Fig. 5A, left, middle). Likewise, feedback projections V1 L2/3 ← LM L5 and V1 L5 ← LM L5 are strengthened as well in the change case (Fig. 5A, left, middle). Even at the scale of cell types, we observe that the change condition leads to an increase in effective connectivity for both inter-areal feedforward and feedback projections (Fig. 5A, right, magenta boxes). Additionally, we note that the changes are predominantly red, i.e., the inferred weights in the change condition are generally higher than those in the case of expected stimuli and conditions being perceived (Fig. 5A, middle). This trend also holds when we compare familiar and novel stimuli (section S8 and fig. S2) in both the change and no change cases, i.e., the introduction of novelty leads to increased inter/intra-area connectivity. In agreement with previous literature (*67*), this suggests that novelty and unexpectedness increase the brain's excitability, which in turn could facilitate plasticity and aid learning.

***Familiar no change versus familiar change (half-set presentation)***

When focusing on the half-set presentation for familiar images, however, we found a contrasting pattern of connectivity with the feedback signaling (Fig. 5B, left, middle). While we see almost no difference in the weights V1 L5 ← LM L5 across the change and no change cases, the weights V1 L2/3 ← LM L5 are distinctly higher in the no change case than the change case. The feedforward projection V1 L2/3 → LM L4, however, still maintains the same trend as the full presentation case, wherein it is higher when an image change occurs than when it does not. This suggests that while the feedforward communication is relatively immediate using shorter timescales, propagating feedback information occurs over a longer timescale (*68*–*70*), making a case for further investigation of role of inter-areal, cortico-cortical time delays (*71*) when studying predictive coding (*72*). The cell type–specific analysis further reveals that Vip neurons in L2/3 are less inhibited by Sst neurons of the same layer in the change case in both V1 and LM (Fig. 5B, right, green boxes), compared to the full presentation case (Fig. 5A, right, green boxes), once again speaking to the transience of the change and also supporting the idea that Vip neurons could encode unexpectedness as hypothesized by the predictive coding theory.

***Familiar change versus familiar omission***

Our setup also allows us to compare how the network processes different types of expectation violations, by contrasting the learnt weights in the case of an image change versus image omission. In the setting of familiar images over the length of a full-set presentation (Fig. 5C), we notice that while most of the weights are quite similar for both types of expectation violations, there is an appreciable increase in the feedback projection V1 L5 ← LM L5 in the case of image omission (Fig. 5C, left, middle). Additionally, this increase seems to be driven by an increase in the feedback connection weight to the Vip cells in V1 L5, as well as an overall increase in the connectivity weights targeted to V1 L5 Vip neurons (Fig. 5C, right). These observations are in agreement with experimental findings that omissions trigger signaling in Vip neurons in V1 (*73*). We also note that across the RNN, weights to the Vip neurons from locally adjacent Sst neurons are reduced in the omission case, suggesting that these neurons are not as inhibited during the processing of omissions, thus potentially emphasizing their role in processing prediction violations. In the Supplementary Materials, we also provide results studying the no change versus omission case with both familiar and novel images (section S8 and fig. S3) for fair comparison.

Collectively, our analysis demonstrates a hierarchical organization of predictive processing in the visual cortex operating over different timescales. We find that feedforward projections are consistently enhanced during all prediction violations across both long and short timescales, emphasizing their crucial role in transmitting prediction error signals [and fundamentally driving synaptic plasticity in the brain (*74*–*76*), facilitating learning and adaptation]. In contrast, feedback projections are modulated by both the type of prediction error and the temporal window over which neuronal responses are modeled. Notably, the behavior of feedback projections differs when targeting different cortical layers: Feedback projections to L2/3 are more prominently modulated during unviolated predictions over shorter timescales (Fig. 5B), while feedback projections to L5 are more responsive during negative prediction errors such as omissions of expected visual input (*77*) (Fig. 5C). This differential modulation suggests that while feedforward pathways rapidly convey unexpected sensory information, feedback pathways adjust more selectively based on the context, timing, and targeted cortical layer of the prediction error. These patterns are further corroborated by our observations comparing change, no change, and omission across familiar

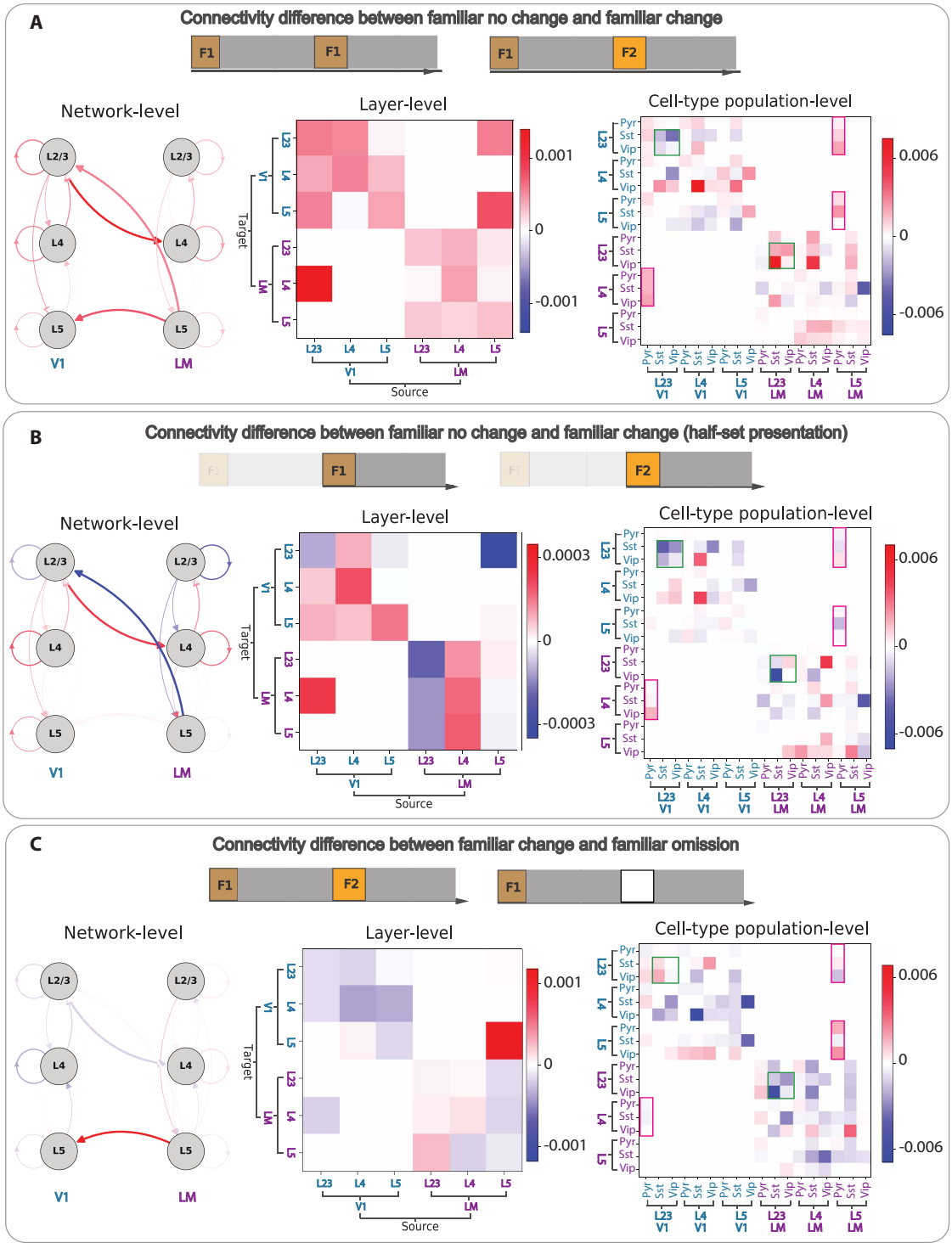


**Fig. 5. Connectivity differences across timescales and test conditions.** (**A**) Familiar no change versus familiar change (full-set presentation). (**B**) Familiar no change versus familiar change (half-set presentation). (**C**) Familiar change versus familiar omission (full-set presentation). All difference connection weights are computed as Second condition − First condition and averaged across individual neuron-to-neuron connections within each population; blue indicates higher first-condition weights, and red indicates higher second-condition weights. Left and middle plots show weights averaged across layers, and right plots show weights averaged by cell type within each layer. Magenta boxes highlight inter-areal feedforward and feedback connections. Green boxes highlight all Sst-ViP interactions in L2/3 of both V1 and LM.

and novel conditions during the full-set presentation (section S8 and fig. S2). In particular, we note that the presentation of a novel image always increases the feedforward connectivity (and ergo the projection) from V1 L2/3 → LM L4. On the finer-scale level of cell types instead of entire layers, there is consistent involvement of Vip interneurons during prediction violations, suggesting that these neurons may play a role in modulating cortical circuits in response to unexpected stimuli.

Overall, our findings illustrate how the brain dynamically adjusts its functional neural connectivity in response to varying predictive contexts, timescales, and cortical layers, providing circuit-level observations that are compatible with the predictive coding framework. The dynamic interplay of feedforward and feedback mechanisms facilitates efficient processing of sensory information, enabling the brain to anticipate and adapt to constantly changing environments. We further note that these directions of inferred connectivity are not typically respected neither when we train without the sign constraints enforced by different cell types (section S9 and fig. S4), i.e., in violation of Dale's law, nor when we train with shuffled interpopulation dynamics, i.e., while keeping the overall structure but having the "physical" L2/3, L4, and L5 fit to the responses of L4, L5, and L2/3, respectively, instead in both areas V1 and LM (section S9 and fig. S5). Results for all comparisons across both the full-set and half-set presentations are publicly available on the project GitHub at: https://github.com/HChoiLab/biologicalRNNs/tree/main/celltype-comparisons.

## DISCUSSION

Our work develops methods for constructing RNNs that simultaneously incorporate two fundamental biological constraints: Dale's law and structured sparse connectivity motifs. We provide mathematical grounding for these methods, including convergence guarantees and error bounds, demonstrating that they can match the performance of unconstrained RNNs. Empirical results on standard synthetic tasks support the efficacy of our approach, demonstrating that our biologically constrained RNNs can achieve performance comparable to conventional, unconstrained networks. Furthermore, by aligning computational models more closely with biological reality, we enhance their utility for neuroscientific research, providing tools for more accurate modeling of neural dynamics and brain function.

Our approach also differs appreciably from CURBD (*20*), an existing method in the literature for inferring multi-regional interactions, in two key aspects. First, while CURBD successfully models neural dynamics and interactions, it does not incorporate sign constraints during training, limiting its ability to differentiate between excitatory and inhibitory cellular mechanisms. Second, and more critically, CURBD's reliance on FORCE training makes it less suitable for implementing experimentally informed sparse connectivity patterns among neuronal populations. Every iteration with FORCE is a least-squares update that is dense and does not respect the sparsity constraints of the matrix at the previous iteration—it is nontrivial to subsequently enforce the sparsity pattern, or alternatively solve a recursive least-squares update for every sub-matrix defined by the sparsity pattern at each update, which quickly becomes computationally infeasible. These limitations consequently motivated our development of a backpropagation-based weight update method that efficiently handles both Dale's law constraints and structured sparsity.

Applying our methods to the Allen Institute Visual Behavior dataset, we inferred multi-regional neuronal interactions underlying visual behavior in mice performing a change detection task. Our anatomically and physiologically constrained CelltypeRNNs not only replicated the experimental data but also provided insights consistent with the theory of predictive coding. Specifically, the models revealed dynamic interplay between feedforward and feedback mechanisms across cortical layers and cell types, capturing how the brain adjusts functional neural connectivity in response to varying predictive contexts and timescales.

We note that much of our methodological work can easily be extended to other deep architectures and is not restricted to simply RNNs. That said, a key area for incorporation of additional biological realism would be in the way we inherently solve the credit assignment problem. Backpropagation suffers from needing a global error signal and weight symmetry (*78*), prompting the need for more biologically plausible learning rules that can still learn as effectively. One hypothesis is that using local learning rules may contribute to the emergence of more modular network representations by promoting the formation of localized activity clusters, thus leading to deeper insights into how functional specialization arises in neural systems and its role in facilitating learning.

Furthermore, our findings highlight differential neural responses to different types of prediction errors, emphasizing the importance of the nature of the violations in shaping neural dynamics. In our study, the change in the familiar image case represents a "global oddball"—an unexpected stimulus that violates established patterns while maintaining the local context. Conversely, the omission of an expected stimulus constitutes a "local oddball," introducing a novel scenario for the network. This distinction is notable, as recent work (*79*) has found that global oddballs elicit responses in nongranular layers, differing from local oddballs that evoke early responses in superficial layers 2/3, consistent with conventional predictive coding theory. Our findings align with this pattern for global oddballs but present discrepancies in the case of local oddballs (omissions). This underscores the need for further exploration into stimulus dependency in error encoding (*80*) and suggests that normative predictive coding computations may need to account for the type of prediction error to fully capture neural processing dynamics.

Finally, we note that an important consideration in our study is the limited scope of recorded cell types and brain regions, which poses challenges in interpreting our results. Specifically, we do not have recordings from all interacting cell types and areas that may be involved in the visual processing tasks we modeled. This limitation means that our models might capture neural responses that are more correlational rather than causal, as they are based solely on the observed data from recorded populations. The absence of data could lead to incomplete or biased representations of neural interactions, especially at the finer grained level of cell types as opposed to the coarser level of layers, where the absence of a particular subpopulation's influence is more easily subsumed within aggregate dynamics. To address this gap, future work could involve developing methods that account for unobserved interactions, perhaps through incorporating previous knowledge of anatomical and effective connectivity or using computational techniques to infer missing information. Additionally, expanding experimental recordings to include more brain areas and cell types would provide a more comprehensive

dataset, enabling our models to capture the full complexity of neural dynamics and leading to more causally robust conclusions.

## MATERIALS AND METHODS

### Dale's backpropagation: Technical details

Consider the typical Elman RNN (*81*) whose hidden states $h_t$ at time $t$ are updated as per the rule

$$\mathbf{h}_t = \phi\left(\mathbf{W}_{hi}\mathbf{x}_t + \mathbf{b}_{hi} + \mathbf{W}_{hh}\mathbf{h}_{t-1} + \mathbf{b}_{hh}\right) \tag{3}$$

where $\phi$ is the nonlinear activation function, $\mathbf{W}_{hh}$ is the recurrent weight matrix, and $\mathbf{W}_{hi}$ is the projection matrix that acts on inputs $x_t$. Biases corresponding to the input and hidden states are denoted as $\mathbf{b}_{hi}$, $\mathbf{b}_{hh}$, respectively.

When $\mathbf{h}_t$ are non-negative, respecting Dale's law simplifies to constraining the recurrent weights $\mathbf{W}$ such that if $i$ is the presynaptic neuron and $j$ is the post-synaptic neuron

$$\mathbf{W} = \begin{cases} \mathbf{W}_{ji} \geq 0 \text{ if neuron } i \text{ is excitatory} \\ \mathbf{W}_{ji} \leq 0 \text{ if neuron } i \text{ is inhibitory} \end{cases}$$

At initialization, the recurrent matrix $\mathbf{W}$ can satisfy the sign constraints by construction. However, given that standard gradient descent-based backpropagation update

$$\mathbf{W}^{(i+1)} = \mathbf{W}^{(i)} - \eta \nabla \ell\left(\mathbf{W}^{(i)}\right) \tag{4}$$

with the step size $\eta$ and loss function $\ell$, there is no guarantee that the updated weights $\mathbf{W}^{(i+1)}$ at the next iteration will satisfy the sign constraints set by Dale's law even if they are respected by the matrix $\mathbf{W}^{(i)}$ at iteration $i$.

We note, however, that our sign constraints always form a convex set (*82*), enabling us to adapt any gradient-based optimization scheme (e.g., SGD, ADAM, RMSprop, etc.) into its projected version (*83*). Hence, after the standard backprop update at every iteration, we project the weights onto their feasible set—the orthant in parameter space where the sign constraints of all individual synaptic weights are met. Mathematically, this projection-based update rule can be expressed as

$$\begin{aligned} \mathbf{W}_{\mathrm{D}}^{(i)} &= \mathcal{P}_C\left(\mathbf{W}^{(i)}\right) \\ &= \mathcal{P}_C\left[\mathbf{W}_{\mathrm{D}}^{(i-1)} - \eta \nabla \ell\left(\mathbf{W}_{\mathrm{D}}\right)^{(i-1)}\right] \\ &= \max\left(0, \mathbf{W}_{[N^+]}^{(i)}\right) \oplus \min\left(0, \mathbf{W}_{[N^-]}^{(i)}\right) \end{aligned} \tag{5}$$

where $\mathbf{W}_{\mathrm{D}}^{(i)}$ represents the weight matrix that satisfies Dale's law at iteration $i$ and $\mathcal{P}_C$ denotes the projection operator. Weight subsets corresponding to excitatory and inhibitory neurons are denoted by $[N^+]$ and $[N^-]$, respectively. The full derivation of the update is provided in section S1.1. The explicit algorithm for the update under gradient descent as the optimizer is shown in section S1.2.

Moreover, this projection onto the feasible set has both a simple interpretation and implementation. At every iteration, weights that violate their assigned sign constraints are set to zero, while those that comply are retained at their updated values. The projection itself can be efficiently implemented by multiplying the weights with a binary mask after each update. This flexibility makes it easy to apply our method across various architectures, seamlessly integrating sign constraints within standard backpropagation frameworks.

Consequently, for a single-layer RNN with $N$ neurons, of which $N^+$ are excitatory and $N^-$ are inhibitory, our entire algorithm for Dale's backpropagation can be summarized as follows:

### Topologically informed probabilistic pruning: Additional details

The top-prob approach is also grounded from an ML and mathematical standpoint. Magnitude-based pruning has a long history (*48*, *84*, *85*) and in its iterative form is still a highly competitive empirical baseline for neural network compression via pruning (*39*, *40*). It also closely relates to methods that look to preserve weights that maintain the dynamics of the network in the spectral sense (*49*, *86*–*88*). Finally, it provides us with an elegant way of maintaining the structural integrity of the network. Recent works have established that the MST of a graph fully encapsulates its zeroth-order topological information (*62*, *89*–*91*). [In particular, for a more thorough treatment of the question why the MST of a graph contains its zeroth-order topological information, we refer the interested reader to sections 2 and 3 of (*91*).] Ergo, probabilistically maintaining higher magnitude weights of the network results in us preserving this topological structure (section S5.1).

While we use top-prob pruning in the one-shot sense, we note, however, that our approach can just as easily be used at initialization to sub-select a sparse network pretraining or alternatively used in the iterative manner that is more typical in the ML community, especially for tasks that are more complicated and less amenable to drastic drops in sparsity levels from a fully trained dense configuration. Additional explanations for how we derive and renormalize our hyperparameter $\kappa$ across different contingencies, as well as adjust the parameter $s$ are provided in sections S4.1, S4.2, and S4.3, respectively.

### Inferring effective connectivity from visual behavior recordings in mice

In the following subsections, we provide details of the specific experimental setup and curated dataset, followed by our model architecture and training methodology.

### Dataset and experimental setup

The Visual Behavior Dataset (*64*, *65*) entails a visually guided, go/no-go task where mice are shown a continuous series of briefly presented natural images and they earn water rewards by correctly reporting when the identity of the image changes (*92*). Responses from the mice are collected as they are presented with two different sets of images: a familiar set (Fig. 4B, top row) comprising images that they were trained on, and a novel set (Fig. 4B, bottom row) that are only presented at test time, during the recordings. While the trials themselves are longitudinal spanning multiple image changes, we restrict ourselves to modeling two full image presentations (Fig. 4C, top). If the identity of the second image is the same as that of the first one, we refer to the condition as no change (Fig. 4C, second row). If the identity of the second image is different from that of the first one, we refer to it as the change condition (Fig. 4C, third row). Both images are always from the same set, i.e., they are both either familiar or novel. In a small subset of the trials (~5%), the second image is omitted, and instead replaced by a blank screen (Fig. 4C, first row), allowing for analysis of expectation signals. We call this the omission condition. For more details on the experimental setup, see (*65*).

1. **Initialize** $W_D^{(0)} = W^{(0)} \in \mathbb{R}^{N \times N}$ such that $W^{(0)} = W_{[N^+]}^{(0)} \oplus W_{[N^-]}^{(0)}$.

   $W_{[N^+]}^{(0)}, W_{[N^-]}^{(0)}$ represent weights from the excitatory and inhibitory neurons respectively.

2. Enforce Dale's law by setting $W_{[N^+]}^{(0)} \in \mathbb{R}_{\geq 0}^{N \times N^+}$ and $W_{[N^-]}^{(0)} \in \mathbb{R}_{\leq 0}^{N \times N^-}$.

3. Sample $W_{[N^+]}^{(0)}$ from $U\left[0, \frac{1}{\sqrt{N}}\right]$ and $W_{[N^-]}^{(0)}$ from $U\left[\frac{-1}{\sqrt{N}}, 0\right]$.

4. Initialize $h_0 \leftarrow \vec{0}$.

5. **For** each time step $t$, compute and threshold the hidden state $h_t$ to be non-negative as:

$$h_t^+ = (\phi(W_{hi}x_t + b_{hi} + W_D h_{t-1}^+ + b_{hh}))^+$$

6. **For** each iteration $i$:

   (a) Compute $W^{(i+1)}$ using standard backpropagation.

   (b) Update weights by setting:

$$W_D^{(i+1)} = \max\left(0, W_{[N^+]}^{(i+1)}\right) \oplus \min\left(0, W_{[N^-]}^{(i+1)}\right)$$

**Algorithm 1. Dale's backpropagation.**

For each of our conditions, we consider two temporal windows. In the full-set presentation (Fig. 4C, indicated at the bottom), we model neural activity across the entire two-image sequence [first image (250 ms), interstimulus interval (500 ms), second presentation/omission (250 ms), and post-stimulus interval (500 ms)], which allows us to capture the sustained dynamics underlying predictive computation across time. In contrast, the half-set presentation (Fig. 4C, indicated at the top) models neural activity following the second presentation/omission, enabling us to isolate the transient neural responses that implement the mechanistic components of prediction and error signaling. This complementary approach provides insights into both the overarching dynamics and the immediate neural interactions that support predictive coding, and gives us the flexibility to infer both long-term and short-term functional interactions.

The complete dataset includes multi-regional two-photon data from two hierarchically adjacent areas, VISp (i.e., primary visual cortex or V1) and VISl (i.e., the lateromedial area or LM). For both areas, we collect recordings at depths roughly corresponding to layers 2/3, 4, and 5 in the cortical column for excitatory, i.e., pyramidal (Pyr) neurons and two types of inhibitory neurons, viz. somatostatin (Sst)– and vasoactive intestinal peptide (Vip)–expressing interneurons (sampling depth distributions provided in section S6). In total, we therefore model the activities of 18 different interacting populations (Fig 4A).

To curate the training data for our RNNs, we compute the neuron-averaged response for every experiment corresponding to each of our individual neuronal populations (e.g., LM L5 Vip) from the Allen Institute Visual Behavior-2P dataset (*64*). We then randomly sample (with replacement) 100 averaged responses from the total set of averaged responses, take their mean, and pass the same through a one-dimensional Gaussian filter ($\sigma = 1$) to produce a single training sample (Fig. 4C, black curve). We subsequently produce 2000 such samples for each of our individual neuronal populations.

### CelltypeRNN: Architecture and training

We model the data as described previously with the anatomically constrained CelltypeRNN that replicates the inter-areal structure of the canonical cortical microcircuit with two hierarchically related cortical areas (Fig. 4A) (*68*, *72*, *93*, *94*) while simultaneously enforcing intra-areal lateral connectivity among different cell types within the cortical column as established by (*66*). Moreover, given that the CelltypeRNN is constructed to be able to replicate experimentally obtained response patterns in different cell populations as specified by their cell type, cortical layer, and area, by learning the connection weights, we in turn represent the inferred functional interactions between the populations across the cortical circuit (*11*, *20*) under different stimulus conditions.

Subsequently, we first train a dense, unbiased Elman RNN using Dale's backpropagation, following which we prune the network's recurrent connections block-wise with top-prob pruning (Fig. 4D, left) to achieve their individual target connectivity sparsities. We subsequently fine-tune the post-pruning nonzero RNN weights to achieve an overall performance that is at least as good as that of the RNN pre-pruning (Fig. 4C, bottom). In our specific instantiation, the ratio of Pyr:Sst:Vip neurons in every layer is 12:2:1 (making the excitatory:inhibitory neuronal ratio 4:1), which with a scaling factor of 16 gives us a total of 240 neurons per layer and 1440 overall in the model. Our lateral connectivity probabilities across populations follow experimental data (*66*) and are explicitly stated in section S7. Longer-range inter-areal projections are sparsified to have a connection probability of 0.3, and are strictly excitatory, i.e., feedforward connections: V1 L2/3 Pyr → LM L4 Pyr, Sst, Vip; feedback connections: V1 L2/3 Pyr, Sst, Vip ← LM L5 Pyr and V1 L5 Pyr, Sst, Vip ← LM L5 Pyr.

In addition to the weights of the RNN—i.e., input weights $W_{hi}$ and recurrent weights $W_{hh}$—we also have readout weights that project the recurrent RNN activity of individual neuronal populations onto their respective output space, using randomly initialized, fully connected linear layers. The readout weights are frozen at the time of initialization of the dense RNN itself and remain so throughout the training procedure. By doing so, we ensure that any changes in the model's behavior come from changes in the recurrent dynamics, and not the model "cheating" by simply adjusting its output mapping. It subsequently also makes it easier to interpret and compare how the internal representations and computations change across conditions. To that end, we also mask the input weight matrix $W_{hi}$ so that recurrent neurons corresponding to a specific population do not receive inputs from any other populations.

Our training objective requires each individual population to be able to reconstruct its activity predictively one time step into the future (Fig. 4C, bottom), giving us the loss function

$$\mathcal{L}_{\text{total}} = \frac{1}{n_{\text{pop}}} \sum_{n=1}^{n_{\text{pop}}} \sum_{t=1}^{T-1} \|\mathbf{x}_{n,t+1} - \hat{\mathbf{x}}_{n,t+1}\|_2^2 \quad (6)$$

where $n_{\text{pop}}$ is the number of interacting neuronal populations and $T$ is the total number of time steps in the sequence. $\mathbf{x}_{n,t+1}$ is the input that will be received for population $n$ at time step $t+1$, while $\hat{\mathbf{x}}_{n,t+1}$ is that predicted by the RNN. The loss function is kept the same during both the dense training (Fig. 4C, bottom, purple curve) and fine-tuning post-pruning stages (Fig. 4C, bottom, green curve). However, we fine-tune for only half the number of epochs (50) as we train for with the dense network (100).

We train separate models for each of our 12 different conditions (familiar/novel × change/no change/omission × full-set/half-set presentation) and compare their connection weights across various spatial scales, the results of which are discussed in the following subsection.

All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Our full codebase to (i) download and preprocess the data, (ii) construct and train the CelltypeRNN models across various conditions and timescales, and (iii) reproduce all figure components is made publicly available at: https://hchoilab.github.io/biologicalRNNs. We note that no animal experiments were carried out on our end and all the data used were obtained from the Allen Institute's publicly available visual behavior dataset (*64*).

## Supplementary Materials

**This PDF file includes:**

Supplementary Text
Figs. S1 to S5
Table S1
References

## REFERENCES AND NOTES


1. Y. Cohen, T. A. Engel, C. Langdon, G. W. Lindsay, T. Ott, M. A. K. Peters, J. M. Shine, V. Breton-Provencher, S. Ramaswamy, Recent advances at the interface of neuroscience and artificial neural networks. *J. Neurosci.* **42**, 8514–8523 (2022).
2. A. Saxe, S. Nelli, C. Summerfield, If deep learning is the answer, what is the question? *Nat. Rev. Neurosci.* **22**, 55–67 (2021).
3. B. A. Richards, T. P. Lillicrap, P. Beaudoin, Y. Bengio, R. Bogacz, A. Christensen, C. Clopath, R. P. Costa, A. de Berker, S. Ganguli, C. J. Gillon, D. Hafner, A. Kepecs, N. Kriegeskorte, P. Latham, G. W. Lindsay, K. D. Miller, R. Naud, C. C. Pack, P. Poirazi, P. Roelfsema, J. Sacramento, A. Saxe, B. Scellier, A. C. Schapiro, W. Senn, G. Wayne, D. Yamins, F. Zenke, J. Zylberberg, D. Therien, K. P. Kording, A deep learning framework for neuroscience. *Nat. Neurosci.* **22**, 1761–1770 (2019).
4. T. C. Kietzmann, P. McClure, N. Kriegeskorte, Deep neural networks in computational neuroscience. *Oxford Res. Encycl. Neurosci.* 10.1093/acrefore/9780190264086.013.46 (2019).
5. D. L. Yamins, J. J. DiCarlo, Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **19**, 356–365 (2016).
6. O. Barak, Recurrent neural networks as versatile tools of neuroscience research. *Curr. Opin. Neurobiol.* **46**, 1–6 (2017).
7. G. R. Yang, X.-J. Wang, Artificial neural networks for neuroscientists: A primer. *Neuron* **107**, 1048–1070 (2020).
8. M. T. Kaufman, M. M. Churchland, S. I. Ryu, K. V. Shenoy, Cortical activity in the null space: Permitting preparation without movement. *Nat. Neurosci.* **17**, 440–448 (2014).
9. M. G. Perich, J. A. Gallego, L. E. Miller, A neural population mechanism for rapid learning. *Neuron* **100**, 964–976.e7 (2018).
10. J. D. Semedo, A. Zandvakili, C. K. Machens, M. Y. Byron, A. Kohn, Cortical areas interact through a communication subspace. *Neuron* **102**, 249–259.e4 (2019).
11. M. G. Perich, K. Rajan, Rethinking brain-wide interactions through multi-region 'network of networks' models. *Curr. Opin. Neurobiol.* **65**, 146–151 (2020).
12. L. Kozachkov, M. Ennis, J.-J. Slotine, RNNs of RNNs: Recursive construction of stable assemblies of recurrent neural networks. *Adv. Neural Inf. Process. Syst.* **35**, 30512–30527 (2022).
13. D. Sussillo, L. F. Abbott, Generating coherent patterns of activity from chaotic neural networks. *Neuron* **63**, 544–557 (2009).
14. B. DePasquale, C. J. Cueva, K. Rajan, G. S. Escola, L. Abbott, full-FORCE: A target-based method for training recurrent networks. *PLOS ONE* **13**, e0191527 (2018).
15. D. Sussillo, M. M. Churchland, M. T. Kaufman, K. V. Shenoy, A neural network that finds a naturalistic solution for the production of muscle activity. *Nat. Neurosci.* **18**, 1025–1033 (2015).
16. V. Mante, D. Sussillo, K. V. Shenoy, W. T. Newsome, Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).
17. C. Pandarinath, D. J. O'Shea, J. Collins, R. Jozefowicz, S. D. Stavisky, J. C. Kao, E. M. Trautmann, M. T. Kaufman, S. I. Ryu, L. R. Hochberg, J. M. Henderson, K. V. Shenoy, L. F. Abbott, D. Sussillo, Inferring single-trial neural population dynamics using sequential auto-encoders. *Nat. Methods* **15**, 805–815 (2018).
18. A. A. Russo, R. Khajeh, S. R. Bittner, S. M. Perkins, J. P. Cunningham, L. F. Abbott, M. M. Churchland, Neural trajectories in the supplementary motor area and motor cortex exhibit distinct geometries, compatible with different classes of computation. *Neuron* **107**, 745–758.e6 (2020).
19. K. Rajan, C. D. Harvey, D. W. Tank, Recurrent network models of sequence generation and memory. *Neuron* **90**, 128–142 (2016).
20. M. G. Perich, C. Arlt, S. Soares, M. E. Young, C. P. Mosher, J. Minxha, E. Carter, U. Rutishauser, P. H. Rudebeck, C. D. Harvey, K. Rajan, Inferring brain-wide interactions using data-constrained recurrent neural network models. bioRxiv 423348 [Preprint] (2020). https://doi.org/10.1101/2020.12.18.423348.
21. N. Maheswaranathan, A. Williams, M. Golub, S. Ganguli, D. Sussillo, Reverse engineering recurrent networks for sentiment classification reveals line attractor dynamics. *Adv. Neural Inf. Process. Syst.* **32** (2019).
22. D. Sussillo, O. Barak, Opening the black box: Low-dimensional dynamics in high-dimensional recurrent neural networks. *Neural Comput.* **25**, 626–649 (2013).

23. D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, J. J. Di Carlo, Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 8619–8624 (2014).
24. A. J. Kell, D. L. Yamins, E. N. Shook, S. V. Norman-Haignere, J. H. McDermott, A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron* **98**, 630–644.e16 (2018).
25. J. Kubilius, M. Schrimpf, K. Kar, R. Rajalingham, H. Hong, N. J. Majaj, E. B. Issa, P. Bashivan, J. Prescott-Roy, K. Schmidt, A. Nayebi, D. Bear, D. L. K. Yamins, J. J. Di Carlo, Brain-like object recognition with high-performing shallow recurrent ANNs. *Adv. Neural Inf. Process. Syst.* **32** (2019).
26. M. Schrimpf, J. Kubilius, H. Hong, N. J. Majaj, R. Rajalingham, E. B. Issa, K. Kar, P. Bashivan, J. Prescott-Roy, K. Schmidt, D. L. K. Yamins, James J. Di Carlo, Brain-score: Which artificial neural network for object recognition is most brain-like? bioRxiv 407007 [Preprint] (2018). https://doi.org/10.1101/407007.
27. J. A. Michaels, S. Schaffelhofer, A. Agudelo-Toro, H. Scherberger, A modular neural network model of grasp movement generation. bioRxiv 742189 [Preprint] (2019). https://doi.org/10.1101/742189.
28. A. Nayebi, D. Bear, J. Kubilius, K. Kar, S. Ganguli, D. Sussillo, James J. Di Carlo, D. L. K. Yamins, Task-driven convolutional recurrent models of the visual system. *Adv. Neural Inf. Process. Syst.* **31** (2018).
29. G. W. Lindsay, Convolutional neural networks as a model of the visual system: Past, present, and future. *J. Cogn. Neurosci.* **33**, 2017–2031 (2021).
30. D. Hassabis, D. Kumaran, C. Summerfield, M. Botvinick, Neuroscience-inspired artificial intelligence. *Neuron* **95**, 245–258 (2017).
31. R. Schaeffer, M. Khona, I. Fiete, No free lunch from deep learning in neuroscience: A case study through models of the entorhinal-hippocampal circuit. *Adv. Neural Inf. Process. Syst.* **35**, 16052–16067 (2022).
32. A. H. Marblestone, G. Wayne, K. P. Kording, Toward an integration of deep learning and neuroscience. *Front. Comput. Neurosci.* **10**, 94 (2016).
33. J. C. Eccles, From electrical to chemical transmission in the central nervous system: The closing address of the Sir Henry Dale centennial symposium, Cambridge, 19 September 1975. *Notes Rec. R. Soc. Lond.* **30**, 219–230 (1976).
34. H. Eavani, T. D. Satterthwaite, R. Filipovych, R. E. Gur, R. C. Gur, C. Davatzikos, Identifying sparse connectivity patterns in the brain using resting-state fMRI. *Neuroimage* **105**, 286–299 (2015).
35. M. Kaiser, Connectomes: From a sparsity of networks to large-scale databases. *Front. Neuroinform.* **17**, 1170337 (2023).
36. J. K. Lappalainen, F. D. Tschopp, S. Prakhya, M. M. Gill, A. Nern, K. Shinomiya, S.-Y. Takemura, E. Gruntman, J. H. Macke, S. C. Turaga, Connectome-constrained networks predict neural activity across the fly visual system. *Nature* **634**, 1132–1140 (2024).
37. G. Giacopelli, D. Tegolo, E. Spera, M. Migliore, On the structural connectivity of large-scale models of brain networks at cellular level. *Sci. Rep.* **11**, 4345 (2021).
38. J. Cornford, D. Kalajdzievski, M. Leite, A. Lamarquette, D. M. Kullmann, B. Richards, Learning to live with Dale's principle: ANNs with separate excitatory and inhibitory units. *Int. Conf. Learn. Represent.* 1–27 (2021).
39. J. Frankle, M. Carbin, The lottery ticket hypothesis: Finding sparse, trainable neural networks. *Int. Conf. Learn. Represent.* 1–42 (2019).
40. H. Tanaka, D. Kunin, D. L. Yamins, S. Ganguli, Pruning neural networks without any data by iteratively conserving synaptic flow. *Adv. Neural Inf. Process. Syst.* **33**, 6377–6389 (2020).
41. N. Lee, T. Ajanthan, P. H. Torr, SNIP: Single-shot network pruning based on connection sensitivity. arXiv:1810.02340 [cs.CV] (2018).
42. C. Wang, G. Zhang, R. Grosse, Picking winning tickets before training by preserving gradient flow. arXiv:2002.07376 [cs.LG] (2020).
43. S. Han, J. Pool, J. Tran, W. Dally, Learning both weights and connections for efficient neural network. *Adv. Neural Inf. Process. Syst.* **28** (2015).
44. T. Miconi, Biologically plausible learning in recurrent neural networks reproduces neural dynamics observed during cognitive tasks. *eLife* **6**, e20899 (2017).
45. S. Minni, L. Ji-An, T. Moskovitz, G. Lindsay, K. Miller, M. Dipoppa, G. R. Yang, Understanding the functional and structural differences across excitatory and inhibitory neurons. bioRxiv 680439 [Preprint] (2019). https://doi.org/10.1101/680439.
46. A. Ingrosso, L. Abbott, Training dynamically balanced excitatory-inhibitory networks. *PLOS ONE* **14**, e0220547 (2019).
47. W. Nicola, C. Clopath, Supervised learning in spiking neural networks with FORCE training. *Nat. Commun.* **8**, 2208 (2017).
48. Y. LeCun, J. Denker, S. Solla, Optimal brain damage. *Adv. Neural Inf. Process. Syst.* **2** (1989).
49. E. Moore, R. Chaudhuri, Using noise to probe recurrent neural network structure and prune synapses. *Adv. Neural Inf. Process. Syst.* **33**, 14046–14057 (2020).
50. H. F. Song, G. R. Yang, X.-J. Wang, Training excitatory-inhibitory recurrent neural networks for cognitive tasks: A simple and flexible framework. *PLOS Comput. Biol.* **12**, e1004792 (2016).
51. D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986).
52. L. Zhang, X. Wang, R. Cueto, C. Effi, Y. Zhang, H. Tan, X. Qin, Y. Ji, X. Yang, H. Wang, Biochemical basis and metabolic interplay of redox regulation. *Redox Biol.* **26**, 101284 (2019).
53. C. Tetzlaff, C. Kolodziejski, M. Timme, F. Wörgötter, Synaptic scaling in combination with many generic plasticity mechanisms stabilizes circuit connectivity. *Front. Comput. Neurosci.* **5**, 47 (2011).
54. P. R. Huttenlocher, Synaptic density in human frontal cortex-developmental changes and effects of aging. *Brain Res.* **163**, 195–205 (1979).
55. E. Bullmore, O. Sporns, The economy of brain network organization. *Nat. Rev. Neurosci.* **13**, 336–349 (2012).
56. M. Garrett, S. Manavi, K. Roll, D. R. Ollerenshaw, P. A. Groblewski, N. D. Ponvert, J. T. Kiggins, L. Casal, K. Mace, A. Williford, A. Leon, X. Jia, P. Ledochowitsch, M. A. Buice, W. Wakeman, S. Mihalas, S. R. Olsen, Experience shapes activity dynamics and stimulus coding of VIP inhibitory cells. *eLife* **9**, e50340 (2020).
57. R. P. Rao, D. H. Ballard, Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2**, 79–87 (1999).
58. J. C. Ye, *Geometry of Deep Learning* (Springer, 2022).
59. P. Li, J. Cornford, A. Ghosh, B. Richards, Learning better with Dale's law: A spectral perspective. *Adv. Neural Inf. Process. Syst.* **36** (2024).
60. A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. De Vito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in pytorch. *Adv. Neural Inf. Process. Syst.* **30** (2017).
61. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. De Vito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32** (2019).
62. B. Rieck, M. Togninalli, C. Bock, M. Moor, M. Horn, T. Gumbsch, K. Borgwardt, Neural persistence: A complexity measure for deep neural networks using algebraic topology. arXiv:1812.09764 [cs.LG] (2018).
63. K. Friston, A theory of cortical responses. *Philos. Trans. R. Soc. B Biol. Sci.* **360**, 815–836 (2005).
64. Visual Behavior—2p. https://portal.brain-map.org/circuits-behavior/visual-behavior-2p [accessed 20 November 2024].
65. M. Garrett, P. Groblewski, A. Piet, D. Ollerenshaw, F. Najafi, I. Yavorska, A. Amster, C. Bennett, M. Buice, S. Caldejon, L. Casal, F. D'Orazi, S. Daniel, S. E. J. de Vries, D. Kapner, J. Kiggins, J. Lecoq, P. Ledochowitsch, S. Manavi, N. Mei, C. B. Morrison, S. Naylor, N. Orlova, J. Perkins, N. Ponvert, C. Roll, S. Seid, D. Williams, A. Williford, R. Ahmed, D. Amine, Y. Billeh, C. Bowman, N. Cain, A. Cho, T. Dawe, M. Departee, M. Desoto, D. Feng, S. Gale, E. Gelfand, N. Gradis, C. Grasso, N. Hancock, B. Hu, R. Hytnen, X. Jia, T. Johnson, I. Kato, S. Kivikas, L. Kuan, Q. L'Heureux, S. Lambert, A. Leon, E. Liang, F. Long, K. Mace, I. M. de Abril, C. Mochizuki, C. Nayan, K. North, L. Ng, G. K. Ocker, M. Oliver, P. Rhoads, K. Ronellenfitch, K. Schelonka, J. Sevigny, D. Sullivan, B. Sutton, J. Swapp, T. K. Nguyen, X. Waughman, J. Wilkes, M. Wang, C. Farrell, W. Wakeman, H. Zeng, J. Phillips, S. Mihalas, A. Arkhipov, C. Koch, S. R. Olsen, Stimulus novelty uncovers coding diversity in survey of visual cortex. bioRxiv 528085 [Preprint] (2023). https://doi.org/10.1101/2023.02.14.528085.
66. L. Campagnola, S. C. Seeman, T. Chartrand, L. Kim, A. Hoggarth, C. Gamlin, S. Ito, J. Trinh, P. Davoudian, C. Radaelli, M. H. Kim, T. Hage, T. Braun, L. Alfiler, J. Andrade, P. Bohn, R. Dalley, A. Henry, S. Kebede, A. Mukora, D. Sandman, G. Williams, R. Larsen, C. Teeter, T. L. Daigle, K. Berry, N. Dotson, R. Enstrom, M. Gorham, M. Hupp, S. Dingman Lee, K. Ngo, P. R. Nicovich, L. Potekhina, S. Ransford, A. Gary, J. Goldy, D. McMillen, T. Pham, M. Tieu, L. Siverts, M. Walker, C. Farrell, M. Schroedter, C. Slaughterbeck, C. Cobb, R. Ellenbogen, R. P. Gwinn, C. D. Keene, A. L. Ko, J. G. Ojemann, D. L. Silbergeld, D. Carey, T. Casper, K. Crichton, M. Clark, N. Dee, L. Ellingwood, J. Gloe, M. Kroll, J. Sulc, H. Tung, K. Wadhwani, K. Brouner, T. Egdorf, M. Maxwell, M. McGraw, C. A. Pom, A. Ruiz, J. Bomben, D. Feng, N. Hejazinia, S. Shi, A. Szafer, W. Wakeman, J. Phillips, A. Bernard, L. Esposito, F. D. D'Orazi, S. Sunkin, K. Smith, B. Tasic, A. Arkhipov, S. Sorensen, E. Lein, C. Koch, G. Murphy, H. Zeng, T. Jarsky, Local connectivity and synaptic dynamics in mouse and human neocortex. *Science* **375**, eabj5861 (2022).
67. A. Schulz, C. Miehl, M. J. Berry II, J. Gjorgjieva, The generation of cortical novelty responses through inhibitory plasticity. *eLife* **10**, e65309 (2021).
68. A. M. Bastos, W. M. Usrey, R. A. Adams, G. R. Mangun, P. Fries, K. J. Friston, Canonical microcircuits for predictive coding. *Neuron* **76**, 695–711 (2012).
69. C. A. Bosman, J. M. Schoffelen, N. Brunet, R. Oostenveld, A. M. Bastos, T. Womelsdorf, B. Rubehn, T. Stieglitz, P. de Weerd, P. Fries, Attentional stimulus selection through selective synchronization between monkey visual areas. *Neuron* **75**, 875–888 (2012).
70. J. D. Semedo, A. I. Jasper, A. Zandvakili, A. Krishna, A. Aschner, C. K. Machens, A. Kohn, B. M. Yu, Feedforward and feedback interactions between visual cortical areas use different population activity patterns. *Nat. Commun.* **13**, 1099 (2022).
71. J.-Y. Moon, K. Müsch, C. E. Schroeder, T. A. Valiante, C. J. Honey, Inter-regional delays fluctuate in the human cerebral cortex. *eLife* **13**, RP92459 (2024).
72. A. Balwani, S. Cho, H. Choi, Exploring the architectural biases of the cortical microcircuit. *Neural Comput.* **37**, 1551–1599 (2025).

73. F. Najafi, S. Russo, J. Lecoq, Unexpected events trigger task-independent signaling in VIP and excitatory neurons of mouse visual cortex. *iScience* **28**, 111728 (2025).
74. L. Hertäg, H. Sprekeler, Learning prediction error neurons in a canonical interneuron circuit. *eLife* **9**, e57541 (2020).
75. J. Haarsma, P. C. Fletcher, J. D. Griffin, H. J. Taverne, H. Ziauddeen, T. J. Spencer, C. Miller, T. Katthagen, I. Goodyer, K. M. J. Diederen, G. K. Murray, Precision weighting of cortical unsigned prediction error signals benefits learning, is mediated by dopamine, and is impaired in psychosis. *Mol. Psychiatry* **26**, 5320–5333 (2021).
76. C. K. Starkweather, N. Uchida, Dopamine signals as temporal difference errors: Recent advances. *Curr. Opin. Neurobiol.* **67**, 95–105 (2021).
77. L. Hertäg, C. Clopath, Prediction-error neurons in circuits with multiple neuron types: Formation, refinement, and functional implications. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2115699119 (2022).
78. B. A. Richards, T. P. Lillicrap, Dendritic solutions to the credit assignment problem. *Curr. Opin. Neurobiol.* **54**, 28–36 (2019).
79. J. A. Westerberg, Y. S. Xiong, E. Sennesh, H. Nejat, D. Ricci, S. Durand, B. Hardcastle, H. Cabasco, H. Belski, A. Bawany, R. Gillis, H. Loeffler, C. R. Peene, W. Han, K. Nguyen, V. Ha, T. Johnson, C. Grasso, A. Young, J. Swapp, B. Ouellette, S. Caldejon, A. Williford, P. A. Groblewski, S. R. Olsen, C. Kiselycznyk, C. Koch, J. A. Lecoq, A. Maier, A. M. Bastos, Stimulus history, not expectation, drives sensory prediction errors in mammalian cortex. bioRxiv 616378 [Preprint] (2024). https://doi.org/10.1101/2024.10.02.616378.
80. S. Furutachi, A. D. Franklin, A. M. Aldea, T. Mrsic-Flogel, S. B. Hofer, Cooperative thalamocortical circuit mechanism for sensory prediction errors. *Adv. Neural Inf. Process. Syst.* **633**, 398–406 (2024).
81. J. L. Elman, Finding structure in time. *Cognit. Sci.* **14**, 179–211 (1990).
82. S. P. Boyd, L. Vandenberghe, *Convex Optimization* (Cambridge Univ. Press, 2004).
83. D. P. Bertsekas, *Nonlinear Programming* (Athena Scientific, ed. 3, 2016).
84. M. C. Mozer, P. Smolensky, Skeletonization: A technique for trimming the fat from a network via relevance assessment. *Adv. Neural Inf. Process. Syst.* **1**, 107–115 (1988).
85. S. Hanson, L. Pratt, Comparing biases for minimal network construction with back-propagation. *Adv. Neural Inf. Process. Syst.* **1**, 177–185 (1988).
86. D. A. Spielman, N. Srivastava, "Graph sparsification by effective resistances" in *Proceedings of the Fortieth Annual ACM Symposium on Theory of computing* (ACM, 2008), pp. 563–568.
87. D. A. Spielman, S.-H. Teng, Spectral sparsification of graphs. *SIAM J. Comput.* **40**, 981–1025 (2011).
88. J. Batson, D. A. Spielman, N. Srivastava, S.-H. Teng, Spectral sparsification of graphs: Theory and algorithms. *Commun. ACM* **56**, 87–94 (2013).
89. H. Doraiswamy, J. Tierny, P. J. Silva, L. G. Nonato, C. Silva, Topomap: A 0-dimensional homology preserving projection of high-dimensional data. *IEEE Trans. Vis. Comput. Graph.* **27**, 561–571 (2020).
90. T. Lacombe, Y. Ike, M. Carriere, F. Chazal, M. Glisse, Y. Umeda, Topological uncertainty: Monitoring trained neural networks through persistence of activation graphs. arXiv:2105.04404 [stat.ML] (2021).
91. A. Balwani, J. Krzyston, "Zeroth-order topological insights into iterative magnitude pruning" in *Topological, Algebraic and Geometric Learning Workshops 2022* (PMLR, 2022), pp. 6–16.
92. P. A. Groblewski, D. R. Ollerenshaw, J. T. Kiggins, M. E. Garrett, C. Mochizuki, L. Casal, S. Cross, K. Mace, J. Swapp, S. Manavi, D. Williams, S. Mihalas, S. R. Olsen, Characterization of learning, motivation, and visual perception in five transgenic mouse lines expressing GCaMP in distinct cell populations. *Front. Behav. Neurosci.* **14**, 104 (2020).
93. R. J. Douglas, K. A. Martin, D. Whitteridge, A canonical microcircuit for neocortex. *Neural Comput.* **1**, 480–488 (1989).
94. V. B. Mountcastle, The columnar organization of the neocortex. *Brain* **120**, 701–722 (1997).
95. R. Schneider, *Convex Bodies: The Brunn–Minkowski Theory*, vol. 151 (Cambridge Univ. Press, 2013).
96. H. Kim, G. Papamakarios, A. Mnih, "The Lipschitz constant of self-attention" in *International Conference on Machine Learning* (PMLR, 2021), pp. 5562–5571.
97. H. Federer, *Geometric Measure Theory* (Springer, 2014).


**Acknowledgments:** We thank A. Wu for insightful comments and feedback. **Funding:** This work was supported by the Alfred P. Sloan Foundation Fellowships in Neuroscience (to H.C.) and the National Eye Institute of the National Institutes of Health under Award Number R00 EY030840 (to H.C.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. **Author contributions:** A.B., H.C., and F.N. conceived the project and designed the experiments. A.B. and H.C. developed the methods and theoretical results. A.B. and A.Q.W. conducted the experiments and produced figures for the manuscript with support from H.C. A.B. wrote the original manuscript. A.B., A.Q.W., and H.C. edited and revised the text. A.B., A.Q.W., F.N., and H.C. analyzed the results and reviewed the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. The data are made publicly available by the Allen Institute at: https://portal.brain-map.org/circuits-behavior/visual-behavior-2p. Code to download the data, curate datasets, construct and train models, as well as reproduce figures are all made available on Zenodo (https://doi.org/10.5281/zenodo.15620871). A maintained version of the code is available on GitHub (https://hchoilab.github.io/biologicalRNNs).

Submitted 4 February 2025
Accepted 11 November 2025
Published 12 December 2025
10.1126/sciadv.adw4970

ScienceAdvances

# Constructing biologically constrained RNNs via Dale's backpropagation and topologically informed pruning

Aishwarya Balwani, Alex Q. Wang, Farzaneh Najafi, and Hannah Choi

*Sci. Adv.* **11** (50), eadw4970. DOI: 10.1126/sciadv.adw4970



*Science Advances* (ISSN 2375-2548) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Advances* is a registered trademark of AAAS.

Science Advances
AAAS

# Supplementary Materials for

## Constructing biologically constrained RNNs via Dale's backpropagation and topologically informed pruning

Aishwarya Balwani *et al.*

Corresponding author: Aishwarya Balwani, abalwani6@gatech.edu; Hannah Choi, hannahch@gatech.edu

*Sci. Adv.* **11**, eadw4970 (2025)
DOI: 10.1126/sciadv.adw4970

**This PDF file includes:**

# Supplementary Text

# 1 Dale's backpropagation update

This section studies in detail the Dale's backpropagation update rule, beginning with the explicit derivation of the same. The following subsections detail the algorithm for implementing this update in a gradient descent framework, and provide proofs regarding the optimality of the resulting weight matrix projection under the Frobenius norm.

## 1.1 Derivation

$$
\begin{aligned}
W_D^{(i+1)} &= \mathcal{P}_C\left(W^{(i+1)}\right) \\
&= \mathcal{P}_C\left(W_D^{(i)} - \eta\nabla\ell\left(W_D^{(i)}\right)\right) \\
&= \mathcal{P}_C\left(W_{D[N^+]}^{(i)} - \eta\nabla\ell\left(W_{D[N^+]}^{(i)}\right)\right) \oplus \mathcal{P}_C\left(W_{D[N^-]}^{(i)} - \eta\nabla\ell\left(W_{D[N^-]}^{(i)}\right)\right) \\
&= \max\left(0, W_{D[N^+]}^{(i)} - \eta\nabla\ell\left(W_{D[N^+]}^{(i)}\right)\right) \oplus \min\left(0, W_{D[N^-]}^{(i)} - \eta\nabla\ell\left(W_{D[N^-]}^{(i)}\right)\right) \\
&= \max\left(0, W_{[N^+]}^{(i+1)}\right) \oplus \min\left(0, W_{[N^-]}^{(i+1)}\right)
\end{aligned}
$$

## 1.2 Algorithm

**Algorithm 2:** Dale's Backpropagation Update Rule (under the gradient descent optimization scheme)

**Input:** Initial weights $W_D^{(0)}$, step size $\eta$, maximum iterations $K$

**Output:** Final weights $W_D^{(K)}$

**for** $k = 0$ **to** $K - 1$ **do**

  Compute gradient $\nabla \ell \left( W_D^{(k)} \right)$;

  $W^{(k+1)} \leftarrow W_D^{(k)} - \eta \nabla \ell \left( W_D^{(k)} \right)$ ; `// Compute weight updates with backpropagation`

  **for** *each component* $j$ **do**

    **if** $sign \left( W_j^{(k+1)} \right) = sign \left( W_{D_j}^{(k)} \right)$ **then**

      $W_{D_j}^{(k+1)} \leftarrow W_j^{(k+1)}$ ; `// Keep weight update if sign constraint is respected`

    **else**

      $W_{D_j}^{(k+1)} \leftarrow 0$ ; `// Set weight to 0 if sign constraint is violated`

    **end**

  **end**

**end**

**return** $W_D^{(K)}$

## 1.3 Closest sign-constrained projection under the Frobenius norm

**Theorem 3** (Dale's backpropagation provides the closest sign-constrained projection of $W$ under the Frobenius norm)**.** *Let $W \in \mathbb{R}^{N \times N}$ be a real square matrix. Define the set $S \subset \mathbb{R}^{N \times N}$ as, (i) Columns 1 to k: All entries are non-negative ($\geq 0$), (ii) Columns $k + 1$ to N: All entries are non-positive ($\leq 0$). Then, the matrix $W_D$ obtained by applying the Dale's backprop update to W is the closest projection of W onto S under the Frobenius norm. That is,*

$$W_D = \arg\min_{X \in S} \|W - X\|_F.$$

*Proof.* To find the projection of $W$ onto $S$ under the Frobenius norm, we need to solve:

$$\min_{X \in S} \|W - X\|_F^2 = \min_{X \in S} \sum_{i=1}^{N} \sum_{j=1}^{N} (W_{ji} - X_{ji})^2.$$

Since the Frobenius norm is separable over the entries of $W$ and $X$, we can minimize each $(W_{ji} - X_{ij})^2$ independently, subject to the sign constraints on $X_{ji}$:

- **For columns** 1 **to** $k$**:** The constraint is $X_{ji} \geq 0$.
- **For columns** $k+1$ **to** $N$**:** The constraint is $X_{ji} \leq 0$.

**Minimization for each** $X_{ji}$**:**

- **If** $j \leq k$**:**

$$X_{ij}^* = \arg\min_{x \geq 0} (W_{ji} - x)^2.$$

  The solution is:

$$X_{ji}^* = \begin{cases} W_{ji}, & \text{if } W_{ji} \geq 0, \\ 0, & \text{if } W_{ji} < 0. \end{cases}$$

- **If** $j > k$**:**

$$X_{ji}^* = \arg\min_{x \leq 0} (W_{ji} - x)^2.$$

  The solution is:

$$X_{ji}^* = \begin{cases} W_{ji}, & \text{if } W_{ji} \leq 0, \\ 0, & \text{if } W_{ij} > 0. \end{cases}$$

Therefore, the optimal $X_{ij}^*$ corresponds exactly to the entries of $W_D$ obtained by Dale's backprop and $W_D$ is the projection of $W$ onto $S$ under the Frobenius norm. □

**Corollary 4.** *Let $W_R$ be the matrix obtained from $W$ using rectified backprop. Then, the Dale's backprop matrix $W_D$ is closer to $W$ than $W_R$ is to $W$ under the Frobenius norm:*

$$\|W - W_D\|_F \leq \|W - W_R\|_F.$$

*Proof.* By Theorem 3, $W_D$ is the closest matrix in $S$ to $W$ under the Frobenius norm. Since $W_R \in S$, it follows that:

$$\|W - W_D\|_F \leq \|W - W_R\|_F$$

□

# 2 Alignment of Dale's backpropagation and Hebbian learning in reinforcing high-magnitude weights

We show that despite the differences in their explicit formulations, both Hebbian learning and Dale's backpropagation tend to strengthen (i.e., increase the magnitude of) similar weights. In particular, the weights strengthened by Hebbian learning form a subset of those strengthened via the gradient-based Dale's backprop. Given that weights of higher magnitude inevitably influence the effective connectivity amongst the neurons, preserving weights of higher magnitudes implies preserving those weights which would've been important from a statistical learning perspective as well being biologically relevant.

Learning requires a weight $w_{ji}$ joining pre-synaptic neuron $i$ to post-synaptic neuron $j$ be changed according to the rule

$$w_{ji}^{(k+1)} = w_{ji}^{(k)} + \Delta w_{ji}^{(k)}$$

Hebbian learning postulates the update $\Delta w_{ji}$ is given as

$$\Delta w_{ji_{Hebb}} = \eta \cdot a_i \cdot a_j$$

where $a_i,\ a_j$ are the activations of neurons $i, j$ respectively while $\eta$ is the learning rate.

On the other hand, the backpropagation update (without loss of generality, in the absence of bias terms) is given as

$$\Delta w_{ji_{BP}} = -\eta \cdot \frac{\partial \ell}{\partial w_{ji}} = -\eta \cdot \underbrace{\frac{\partial \ell}{\partial a_j} \cdot \phi' \left( \sum_i w_{ji} a_i \right)}_{\varepsilon_j} \cdot a_i$$

where $\ell$ is the loss function, $\phi$ is the activation function, $\varepsilon_j$ is the error corresponding to neuron $j$ computed using the chain rule, and $a_i$ has the same meaning as before.

In Dale's backpropagation, we constrain all activations to be non-negative through a thresholding operation, and weights are restricted to maintain their assigned signs. Under these constraints, the following statements hold true:

**Analysis for $w_{ji} \geq 0$:** In the case of Hebbian learning, given our construction, if $w_{ji}$ is non-negative, we would need both $a_i, a_j$ to be positive to increase $|w_{ji}|$.

For Dale's backprop, for a weight $w_{ji} \geq 0$ to increase in magnitude, we require that

$$\frac{\partial \ell}{\partial w_{ji}} < 0 \implies \frac{\partial \ell}{\partial a_j} < 0.$$

since $\phi'(\cdot)$ is always non-negative for monotonically-increasing $\phi$ such as ReLU and tanh. This means that as $\ell$ decreases, the neuron $a_j$ contributes positively to reducing the loss.

In turn, as learning progresses and reduces the loss $\ell$, this would lead to an increase in $a_j$ when $a_j > 0$, matching Hebbian learning.

**Analysis for $w_{ji} \leq 0$** : In the case of Hebbian learning, if $w_{ji}$ is non-positive, we would require that the *action* of $a_i \leq 0$ and $a_j \geq 0$ to increase $|w_{ji}|$.

In the case of Dale's backprop, for a weight $w_{ji} \leq 0$ to increase in magnitude, we now require

$$\frac{\partial \ell}{\partial w_{ji}} > 0 \implies \frac{\partial \ell}{\partial a_j} > 0.$$

Since $\frac{\partial \ell}{\partial a_j} > 0$ indicates that increasing $a_j$ would increase the loss, the learning process will instead push to decrease $a_j$. Strengthening a negative weight (making it more negative) lowers $z_j = \sum_i w_{ji} a_i$ when $a_i \geq 0$, thereby reducing $a_j = \phi(z_j)$ and in turn helping to reduce the loss $\ell$.

This correspondence between Dale's backprop and Hebbian learning, facilitated by the non-negative activation constraint, suggests that weights strengthened during learning align with those of biological significance. Consequently, when these weights are preferentially retained by our pruning rule, we preserve functionally important connectivity patterns that emerge through biologically plausible learning dynamics.

# 3 Theoretical guarantees for Dale's backpropagation

## 3.1 Analyzing convergence of Dale's backpropagation under the restricted optimum assumption

**Lemma 5** (Optimal sign pattern preservation)**.** *Let the vector of learnt weights be $W \in \mathbb{R}^n$ with the components $w_j$, where $j \in \{1, 2, ..., n\}$. Let $L$ be the Lipschitz constant for the gradients $\nabla\ell(W)$, where $\ell$ is a loss function. Given a gradient descent-based, component-wise sign-preserving learning rule that uses the projection operator $\mathcal{P}_C : \mathbb{R}^n \mapsto \mathbb{R}^n$ defined as*

$$\mathcal{P}_C(w_j) = \begin{cases} w_j & \text{if } sign(z_j) = sign\left(w_j\right) \\ \\ 0 & \text{if } sign(z_j) \neq sign\left(w_j\right) \end{cases}$$

*where $z_j = w_j - \frac{1}{L}\nabla\ell\left(w_j\right)$, $sign(z_j) = \frac{z_j}{|z_j|}$ for $z_j \neq 0$, and $sign(0) = 0$. If $sign(W^*) = sign\left(W^{(0)}\right)$ where $W^*$ are the set of weights that can achieve the optimal loss on $\ell$, it holds that for any iteration $i$ of regular gradient descent*

$$sign\left(W^{(i)}\right) = sign\left(W^{(0)}\right) = sign\left(W^*\right) \;\; \forall i \in \mathbb{N}, \text{ and } \mathcal{P}_C(w_j) = w_j \text{ for } j \in \{1, 2, ..., n\}.$$

*Proof.* We show by induction that $sign\left(W^{(i)}\right) = sign\left(W^{(0)}\right) = sign\left(W^*\right) \;\; \forall i \in \mathbb{N}$.

Base case ($i = 0$): The statement trivially holds true since $sign\left(W^{(0)}\right) = sign\left(W^*\right)$, by assumption.

Inductive hypothesis: For some iteration $i > 0$, $sign\left(W^{(i)}\right) = sign\left(W^{(0)}\right) = sign\left(W^*\right)$.

To show that for the iteration $i + 1$ it also holds that $sign\left(W^{(i+1)}\right) = sign\left(W^{(0)}\right) = sign\left(W^*\right)$, we consider $z_j$ as defined, which is the $j^{th}$ component of

$$z = W - \frac{1}{L}\nabla\ell(W).$$

By the Lipschitz continuity of the gradient, we have that

$$\left\|\nabla\ell\left(w_j^{(i)}\right) - \nabla\ell\left(w_j^*\right)\right\|_2 \leq L\left\|w_j^{(i)} - w_j^*\right\|_2$$

Since $W^*$ is the optimal set of weights, we know that $\nabla \ell(w_j^*) = 0,\ \forall\ j \in \{1, 2, ..., n\}$. Therefore,

$$\left\|\nabla \ell\left(w_j^{(i)}\right)\right\|_2 \leq L \left\|w_j^{(i)} - w_j^*\right\|_2 \implies \frac{1}{L}\left|\nabla \ell\left(w_j^{(i)}\right)\right| \leq \left|w_j^{(i)} - w_j^*\right|$$

Consider the case where $w_j^{(i)} < w_j^*$.

Here, $w_j^{(i)} - w_j^* < 0 \implies \left|w_j^{(i)} - w_j^*\right| = -w_j^{(i)} + w_j^*$. Furthermore, the gradient descent update moves $w_j^{(i)}$ towards $w_j^*$ by increasing its value, implying that $\nabla \ell\left(w_j^{(i)}\right)$ itself is negative. Consequently,

$$\begin{aligned} z_j^{(i)} &= w_j^{(i)} - \frac{1}{L}\nabla \ell\left(w_j^{(i)}\right) \\ &\leq w_j^{(i)} + (-w_j^{(i)} + w_j^*) \\ &= w_j^* \end{aligned}$$

This leads us to the conclusion that $w_j^{(i)} < z_j^{(i)} < w_j^*$.

Since $sign\left(w_j^{(i)}\right) = sign\left(w_j^*\right)$ by the induction hypothesis, $sign\left(w_j^{(i)}\right) = sign\left(z_j^{(i)}\right) = sign\left(w_j^*\right)$ also holds. As a result, $\mathcal{P}_C\left(z_j^{(i)}\right) = z_j^{(i)}$ and $sign\left(w^{(i+1)}\right) = sign\left(w_j^{(i)}\right) = sign\left(w_j^{(0)}\right) = sign\left(w_j^*\right)$.

The case where $w_j^{(i)} > w_j^*$ follows similarly, with the difference that since the gradient $\nabla \ell\left(w_j^{(i)}\right)$ is positive and $\left|w_j^{(i)} - w_j^*\right| = w_j^{(i)} - w_j^*$, we instead have $w_j^{(i)} > z_j^{(i)} > w_j^*$. This leads to the same results as before, i.e., $\mathcal{P}_C\left(z_j^{(i)}\right) = z_j^{(i)}$ and $sign\left(w^{(i+1)}\right) = sign\left(w_j^{(i)}\right) = sign\left(w_j^{(0)}\right) = sign\left(w_j^*\right)$.

As the choice of the index $j$ was arbitrary, these results holds across all indices and therefore

$$sign\left(W^{(i)}\right) = sign\left(W^{(0)}\right) = sign\left(W^*\right)\ \ \forall i \in \mathbb{N}, \text{ and } \mathcal{P}_C(w_j) = w_j \text{ for } j \in \{1, 2, ..., n\}.$$

□

**Theorem** (Convergence of Dale's Backpropagation)**.** *Let $\ell$ be a loss function satisfying the $\mu-$Polyak-Lojasiewicz condition, with gradients that are L-Lipschitz such that $L \geq \mu > 0$. Consider the sequence of weights $\{W_D^{(i)}\}$ generated according to the Dale's backpropagation update, with a step size of $\frac{1}{L}$. Given an optimal loss $\ell^* = \ell(W^*) = \operatorname{argmin} \ell(W_D)$ where $W^*$ has the same sign*

*pattern as all* $W_D^{(i)}$ *and a specific error* $\varepsilon > 0$*, it holds for the iteration* $i$ *that*

$$\ell(W_D^{(i)}) - \ell^* \leq \varepsilon \text{ when } i \geq \frac{\log\left(\frac{\ell(W_D^{(0)}) - \ell^*}{\varepsilon}\right)}{\log\left(\frac{L}{L-\mu}\right)}$$

*Proof.* By Lemma 6 we note that the function $g(W)$ is convex, when it is defined as

$$g(W) = \frac{L}{2}\|W\|_2^2 - \ell(W)$$

Furthermore, by the first-order equivalence of convexity on $g(W)$, we have

$$g(W') \geq g(W) + \langle \nabla g(W), W' - W \rangle \quad \forall\ W, W'$$

This subsequently implies that

$$\frac{L}{2}\|W'\|_2^2 - \ell(W') \geq -\frac{L}{2}\|W\|_2^2 - \ell(W) + L\langle W', W\rangle - \langle W' - W, \nabla\ell(W)\rangle$$

Rearranging terms, we have

$$\ell(W') \leq \ell(W) + \langle \nabla\ell(W), W' - W\rangle + \frac{L}{2}\|W' - W\|_2^2 \tag{S1}$$

Setting $W' = W_D^{(i+1)}$ and $W = W_D^{(i)}$ in Eq. S1 while using the Dale's backprop update rule, we get

$$\ell\left(W_D^{(i+1)}\right) - \ell\left(W_D^{(i)}\right) \leq \left\langle \nabla\ell\left(W_D^{(i)}\right), W_D^{(i+1)} - W_D^{(i)}\right\rangle + \frac{L}{2}\left\|W_D^{(i+1)} - W_D^{(i)}\right\|_2^2 \tag{S2}$$

Defining $z = W_D^{(i)} - \frac{1}{L}\nabla\ell\left(W_D^{(i)}\right)$ and the projection operator $\mathcal{P}_C$ as before, Eq. S2 can be re-written as

$$\begin{aligned}\ell\left(W_D^{(i+1)}\right) - \ell\left(W_D^{(i)}\right) &\leq \left\langle \nabla\ell\left(W_D^{(i)}\right), \mathcal{P}_C(z) - W_D^{(i)}\right\rangle + \frac{L}{2}\left\|\mathcal{P}_C(z) - W_D^{(i)}\right\|_2^2 \\ &= \underbrace{\left\langle \nabla\ell\left(W_D^{(i)}\right), \mathcal{P}_C(z) - z\right\rangle}_{\text{Term 1}} + \underbrace{\left\langle \nabla\ell\left(W_D^{(i)}\right), z - W_D^{(i)}\right\rangle}_{\text{Term 2}} + \underbrace{\frac{L}{2}\left\|\mathcal{P}_C(z) - W_D^{(i)}\right\|_2^2}_{\text{Term 3}}\end{aligned}$$

By Lemma 5 we note that **Term 1** is always 0 since $\mathcal{P}_C(z) = z$.

Re-substituting $z = W_D^{(i)} - \frac{1}{L}\nabla\ell\left(W_D^{(i)}\right)$ in **Term 2** simplifies it to $-\frac{1}{L}\left\|\nabla\ell\left(W_D^{(i)}\right)\right\|_2^2$

Due to the non-expansive property of metric projections onto convex sets (Theorem 1.2.1 of (*95*)) and the fact that $\mathcal{P}_C\left(W_D^{(i)}\right) = W_D^{(i)}$ it holds for **Term 3** that

$$\left\|\mathcal{P}_C(z) - \mathcal{P}_C\left(W_D^{(i)}\right)\right\|_2 \leq \left\|z - W_D^{(i)}\right\|_2 = \frac{1}{L}\left\|\nabla\ell\left(W_D^{(i)}\right)\right\|_2$$

Combining the three terms, we get the bound

$$\begin{aligned}\ell\left(W_D^{(i+1)}\right) - \ell\left(W_D^{(i)}\right) &\leq -\frac{1}{L}\left\|\nabla\ell\left(W_D^{(i)}\right)\right\|_2^2 + \frac{1}{2L}\left\|\nabla\ell\left(W_D^{(i)}\right)\right\|_2^2 \\ &= -\frac{1}{2L}\left\|\nabla\ell\left(W_D^{(i)}\right)\right\|_2^2\end{aligned}$$

Using the Polyak-Lojasiewicz inequality (Def. 3.1) we get

$$\ell\left(W_D^{(i+1)}\right) - \ell\left(W_D^{(i)}\right) \leq -\frac{\mu}{L}\left(\ell\left(W_D^{(i)}\right) - \ell^*\right)$$

Rearranging and subtracting $\ell^*$ from both sides gives us

$$\ell\left(W_D^{(i+1)}\right) - \ell^* \leq \left(1 - \frac{\mu}{L}\right)\left(\ell\left(W_D^{(i)}\right) - \ell^*\right) \tag{S3}$$

Applying Eq. S3 recursively gives us the result

$$\ell\left(W_D^{(i)}\right) - \ell^* \leq \left(1 - \frac{\mu}{L}\right)^i\left(\ell\left(W_D^{(0)}\right) - \ell^*\right) \tag{S4}$$

Let the error $\varepsilon$ be defined as $\varepsilon = \ell\left(W_D^{(i)}\right) - \ell^*$ thereby simplifying Eq. S4 to

$$\varepsilon \leq \left(1 - \frac{\mu}{L}\right)^i\left(\ell\left(W_D^{(0)}\right) - \ell^*\right)$$

Taking the logarithm on both sides we get

$$\log(\varepsilon) \leq i \cdot \log\left(1 - \frac{\mu}{L}\right) + \log\left(\ell\left(W_D^{(0)}\right) - \ell^*\right)$$

Rearranging the terms finally gives us the bound

$$i \geq \frac{\log\left(\frac{\ell(W_D^{(0)}) - \ell^*}{\varepsilon}\right)}{\log\left(\frac{L}{L-\mu}\right)}$$

for the target error $\ell(W_D^{(i)}) - \ell^* \leq \varepsilon$.

□

**Definition 3.1** (Polyak-Lojasiewicz condition)**.** *A loss function $\ell$ is said to satisfy the Polyak-Lojasiewicz condition if for some $\mu > 0$ it holds that:*

$$\frac{1}{2} \|\nabla \ell(W)\|_2^2 \geq \mu(\ell(W) - \ell^*) \ \forall \ W$$

*where $\ell^* = \text{argmin } \ell(W)$ is the optimal loss attainable.*

**Lemma 6** (Convexity of transformed function, Lemma 11.1 of (*58*))**.** *If the gradient of a loss function $\ell(W)$ is L-Lipschitz, then the transformed function $g$ is convex, where $g : \mathbb{R}^n \mapsto \mathbb{R}$ is defined as*

$$g(W) \coloneqq \frac{L}{2} \|W\|_2^2 - \ell(W)$$

*Proof.* Since $\nabla \ell(W)$ is $L$-Lipschitz, we have

$$\|\nabla \ell(W) - \nabla \ell(W')\|_2 \leq L \|W - W'\|_2 \ \forall \ W, W'$$

By the Cauchy-Schwarz inequality, we then have

$$\langle \nabla \ell(W) - \nabla \ell(W'), W - W' \rangle \leq L \|W - W'\|_2^2 \ \forall \ W, W'$$

Rearranging terms,

$$\begin{aligned} 0 &\leq -\langle \nabla \ell(W) - \nabla \ell(W'), W - W' \rangle + L \|W - W'\|_2^2 \\ &= \langle W - W', L(W - W') - \nabla \ell(W) + \nabla \ell(W') \rangle \end{aligned}$$

Substituting $g(W) = \frac{L}{2} \|W\|_2^2 - \ell(W)$ and $\nabla g(W) = LW - \nabla \ell(W)$, we get

$$0 \leq \langle W - W', \nabla g(W) - \nabla g(W') \rangle \ \forall \ W, W'$$

By the monotonicity of the gradient, $g(W)$ is convex. □

## 3.2 Analyzing Dale's backprop with respect to standard backpropagation

**Lemma 7** (Distance between learnt weights)**.** *Let $W^{(i)}$ and $W_D^{(i)}$ be the weights at iteration i for standard backpropagation and Dale's backpropagation, respectively. Assume the gradients $\nabla \ell(W)$ and $\nabla \ell(W_D)$ are upper bounded in magnitude by G and Lipschitz continuous with constant L. Then,*

*the distance between the two sets of weights at any iteration $i$, denoted as $\|\delta^{(i)}\|_2 = \|W^{(i)} - W_D^{(i)}\|_2$, is bounded by:*

$$\|\delta^{(i)}\|_2 \leq \frac{G}{L}\left((1+\eta L)^i - 1\right)$$

*where $\eta$ is the learning rate.*

*Proof.* Consider the case where the weights of the network are updated using gradient descent as the optimizer. This implies the following update rules at any iteration $i$:

1. Standard backpropagation update: $W^{(i)} = W^{(i-1)} - \eta\nabla\ell(W)^{(i-1)}$
2. Dale's backpropagation update: $W_D^{(i)} = \mathcal{P}_C\left(W_D^{(i-1)} - \eta\nabla\ell(W_D)^{(i-1)}\right)$

Let $\|\delta^{(i)}\|_2 = \|W^{(i)} - W_D^{(i)}\|_2$ be the distance between the two sets of weights at iteration $i$. We can bound this as:

$$\begin{aligned}
\left\|\delta^{(i)}\right\|_2 &= \left\|W^{(i)} - W_D^{(i)}\right\|_2 \\
&= \left\|W^{(i)} - \left(W_D^{(i-1)} - \eta\nabla\ell(W_D)^{(i-1)}\right) + \left(W_D^{(i-1)} - \eta\nabla\ell(W_D)^{(i-1)}\right) - W_D^{(i)}\right\|_2 \\
&\leq \left\|W^{(i)} - \left(W_D^{(i-1)} - \eta\nabla\ell(W_D)^{(i-1)}\right)\right\|_2 + \left\|\left(W_D^{(i-1)} - \eta\nabla\ell(W_D)^{(i-1)}\right) - W_D^{(i)}\right\|_2 \\
&= \underbrace{\left\|W^{(i-1)} - \eta\nabla\ell(W)^{(i-1)} - \left(W_D^{(i-1)} - \eta\nabla\ell(W_D)^{(i-1)}\right)\right\|_2}_{\text{Term 1}} \\
&\quad + \underbrace{\left\|\left(W_D^{(i-1)} - \eta\nabla\ell(W_D)^{(i-1)}\right) - \mathcal{P}_C\left(W_D^{(i-1)} - \eta\nabla\ell(W_D)^{(i-1)}\right)\right\|_2}_{\text{Term 2}}
\end{aligned}$$

We now bound each term separately.

**Bounding Term 1:**

$$\begin{aligned}
&\left\|W^{(i-1)} - \eta\nabla\ell(W)^{(i-1)} - \left(W_D^{(i-1)} - \eta\nabla\ell(W_D)^{(i-1)}\right)\right\|_2 \\
&\leq \left\|W^{(i-1)} - W_D^{(i-1)}\right\|_2 + \left\|\eta\nabla\ell(W_D)^{(i-1)} - \eta\nabla\ell(W)^{(i-1)}\right\|_2 \text{ (by Triangle inequality)} \\
&= \left\|\delta^{(i-1)}\right\|_2 + \eta\left\|\nabla\ell(W_D)^{(i-1)} - \nabla\ell(W)^{(i-1)}\right\|_2 \\
&\leq \left\|\delta^{(i-1)}\right\|_2 + \eta L\left\|W_D^{(i-1)} - W^{(i-1)}\right\|_2 \quad \text{(by Lipschitz continuity)} \\
&= \left\|\delta^{(i-1)}\right\|_2 + \eta L\left\|\delta^{(i-1)}\right\|_2 \\
&= \left\|\delta^{(i-1)}\right\|_2 (1+\eta L)
\end{aligned}$$

**Bounding Term 2:** The difference between the update using gradient descent before and after the projection step $\mathcal{P}_C$ at iteration $(i-1)$ will never exceed $\eta\nabla\ell(W_D)^{(i-1)}$ when the update pushes weights $W_D$ outside the feasible region. Therefore,

$$\left\|\left(W_D^{(i-1)} - \eta\nabla\ell(W_D)^{(i-1)}\right) - \mathcal{P}_C\left(W_D^{(i-1)} - \eta\nabla\ell(W_D)^{(i-1)}\right)\right\|_2 \leq \left\|\eta\nabla\ell(W_D)^{(i-1)}\right\|_2 \leq \eta G$$

Combining the two bounds, we get:

$$\left\|\delta^{(i)}\right\|_2 \leq \left\|\delta^{(i-1)}\right\|_2 (1+\eta L) + \eta G$$

This forms a recurrence relation as follows

$$\begin{aligned}
\left\|\delta^{(1)}\right\|_2 &\leq \eta G \\
\left\|\delta^{(2)}\right\|_2 &\leq \eta G(1+\eta L) + \eta G \\
\left\|\delta^{(3)}\right\|_2 &\leq \eta G(1+\eta L)^2 + \eta G(1+\eta L) + \eta G \\
&\vdots
\end{aligned}$$

More generally, for any iteration $i$,

$$\left\|\delta^{(i)}\right\|_2 \leq \eta G \sum_{k=0}^{i-1}(1+\eta L)^k$$

This is a geometric series with ratio $(1+\eta L)$ and $i$ terms. Summing the series we get

$$\left\|\delta^{(i)}\right\|_2 \leq \eta G \cdot \frac{(1+\eta L)^i - 1}{(1+\eta L) - 1} = \frac{G}{L}\left((1+\eta L)^i - 1\right)$$

□

**Theorem** (Differences in errors between solutions)**.** *Let $f(W)$ be the function represented by a single-layer RNN unrolled over $T$ timesteps, with weights $W$. Let $W_D$ be the weights learnt using Dale's backpropagation, and $W$ be the weights learnt using standard backpropagation. Assume the non-linearity $\phi$ is either tanh or ReLU. Then, the error of the solution found using Dale's backpropagation with respect to the ground truth $y$ is bounded by:*

$$\|f(W_D) - y\|_2^2 \leq \delta^2 \sum_{t=1}^{T}(L_{f_t})^2 + \sum_{t=1}^{T}(\varepsilon_t^*)^2$$

*where $\delta = \frac{G}{L}\left((1+\eta L)^K - 1\right)$ after $K$ training iterations, $L_{f_t} = \max\left(L_{f_t(W)}, L_{f_t(W_D)}\right)$ is the Lipschitz constant of the RNN at timestep $t$, and $\varepsilon_t^* = \|f_t(W) - y_t\|_2$ is the error of the solution found using conventional backpropagation at timestep $t$.*

*Proof.* We begin by considering a single-layer RNN unrolled over $T$ timesteps. Let $f(W)$ be the function represented by this network, where $W$ are the weights. We can express $f(W)$ as a composition of functions for each timestep:

$$f(W, x) = (f_T \circ f_{T-1} \circ \cdots \circ f_2 \circ f_1)(x_t) \tag{S5}$$

where each $f_t(W_t, x_t) = \phi(W_{hh}h_{t-1} + W_{hi}x_t)$ represents the function at timestep $t$.

Now, let's consider the Lipschitz constants of these functions. When the non-linearity $\phi$ is either tanh or ReLU (both of which are globally Lipschitz with $L_\phi = 1$), it holds by Lemmas 8 and 9 that for every individual layer $f_t$, we have:

$$L_{f_t(W)} \leq L_\phi \cdot L_{W_t} \tag{S6}$$

By recursively substituting Eq. S6 in Eq. S5, we notice the following pattern

$$h_1 = \phi\left(W_{hh}h_0 + W_{hi}x_1\right)$$

$$L_{f_1(W)} \leq \|W_{hh}\|_2 + \|W_{hi}\|_2$$

$$h_2 = \phi\left(W_{hh}h_1 + W_{hi}x_2\right) = \phi\left(W_{hh}\phi\left(W_{hh}h_0 + W_{hi}x_1\right) + W_{hi}x_2\right)$$

$$\begin{aligned} L_{f_2(W)} &\leq \|W_{hh}\|_2^2 + \|W_{hh}\|_2 \|W_{hi}\|_2 + \|W_{hi}\|_2 \\ &= \|W_{hh}\|_2^2 + \|W_{hi}\|_2 \left(1 + \|W_{hh}\|_2\right) \end{aligned}$$

$$h_3 = \phi\left(W_{hh}h_2 + W_{hi}x_3\right) = \phi\left(W_{hh}\phi\left(W_{hh}\phi\left(W_{hh}h_0 + W_{hi}x_1\right) + W_{hi}x_2\right) + W_{hi}x_3\right)$$

$$\begin{aligned} L_{f_3(W)} &\leq \|W_{hh}\|_2^3 + \|W_{hh}\|_2^2 \|W_{hi}\|_2 + \|W_{hh}\|_2 \|W_{hi}\|_2 + \|W_{hi}\|_2 \\ &= \|W_{hh}\|_2^3 + \|W_{hi}\|_2 \left(1 + \|W_{hh}\|_2 + \|W_{hh}\|_2^2\right) \end{aligned}$$

Generalizing the pattern, the Lipschitz constant $L_{f_t(W)}$ of the RNN at $T$ timesteps is bounded by

$$L_{f_T(W)} \leq \|W_{hh}\|_2^T + \|W_{hi}\|_2 \left(\sum_{t=0}^{T-1} \|W_{hh}\|_2^t\right)$$

Summing the geometric series we get

$$L_{f_T(W)} \leq \begin{cases} \|W_{hh}\|_2^T + \|W_{hi}\|_2 \cdot \left(\frac{1-\|W_{hh}\|_2^T}{1-\|W_{hh}\|_2}\right) & \text{if } \|W_{hh}\|_2 \neq 1 \\ \\ 1 + T \cdot \|W_{hi}\|_2 & \text{if } \|W_{hh}\|_2 = 1 \end{cases} \tag{S7}$$

The Lipschitz constant $L_{f_t(W_D)}$ of the RNN with weights $W_D$ can be bounded similarly.

Now, let's consider the difference between the outputs of the RNNs with weights $W$ and $W_D$ at any given timestep $t$. By Lipschitzness,

$$\begin{aligned} \|f_t(W) - f_t(W_D)\|_2 &\leq \max(L_{f_t(W)}, L_{f_t(W_D)}) \, \|W - W_D\|_2 \\ &= L_{f_t} \, \|W - W_D\|_2 \end{aligned} \tag{S8}$$

where $L_{f_t} = \max\left(L_{f_t(W)}, L_{f_t(W_D)}\right)$.

Applying Lemma 7 after $K$ training iterations, we get:

$$\|W - W_D\|_2 \leq \frac{G}{L}\left((1+\eta L)^K - 1\right) = \delta \tag{S9}$$

Therefore, we can simplify our bound on the difference between the outputs:

$$\|f_t(W) - f_t(W_D)\|_2 \leq L_{f_t} \cdot \delta \tag{S10}$$

Let $y_t$ be the ground truth at timestep $t$. Applying the triangle inequality, we get:

$$\begin{aligned} \|f_t(W_D) - y_t\|_2 &= \|f_t(W_D) - f_t(W) + f_t(W) - y_t\|_2 \\ &\leq \|f_t(W_D) - f_t(W)\|_2 + \|f_t(W) - y_t\|_2 \end{aligned}$$

Let $\varepsilon_t^* = \|f_t(W) - y_t\|_2$ be the error of the solution found using conventional backpropagation at timestep $t$. Hence,

$$\|f_t(W_D) - y_t\|_2 \leq L_{f_t} \cdot \delta + \varepsilon_t^* \tag{S11}$$

To get the overall error, we sum over all timesteps:

$$
\begin{aligned}
\|f(W_D) - y\|_2^2 &= \sum_{t=1}^{T} \|f_t(W_D) - y_t\|_2^2 \\
&\leq \sum_{t=1}^{T} (L_{f_t} \cdot \delta + \varepsilon_t^*)^2 \\
&= \sum_{t=1}^{T} (L_{f_t}^2 \cdot \delta^2 + 2L_{f_t} \cdot \delta \cdot \varepsilon_t^* + (\varepsilon_t^*)^2) \\
&= \delta^2 \sum_{t=1}^{T} L_{f_t}^2 + 2\delta \sum_{t=1}^{T} L_{f_t} \cdot \varepsilon_t^* + \sum_{t=1}^{T} (\varepsilon_t^*)^2 \\
&\leq 2 \cdot \left( \delta^2 \sum_{t=1}^{T} (L_{f_t})^2 + \sum_{t=1}^{T} (\varepsilon_t^*)^2 \right)
\end{aligned}
$$

where the last inequality follows from the fact that $2\delta \sum_{t=1}^{T} L_{f_t} \cdot \varepsilon_t^* \leq \delta^2 \sum_{t=1}^{T} (L_{f_t})^2 + \sum_{t=1}^{T} (\varepsilon_t^*)^2$ over $\mathbb{R}$. □

**Lemma 8** (Lipschitz constant of matrix multiplication (*96, 97*))**.** *For a linear transformation $f(x) = Wx$, the Lipschitz constant $L_f$ is equal to the operator norm of $W$, i.e., $L_f = ||W||_{op}$.*

**Lemma 9** (Lipschitzness of composable Lipschitz functions (*96, 97*))**.** *Let $g$ and $h$ be two composable Lipschitz functions with constants $L_g, L_h$ respectively. Then $g \circ h$ is also Lipschitz with the constant $L_{(g \circ h)} \leq L_g \cdot L_h$.*

# 4 Setting $\kappa$ and sparsity values for pruning

## 4.1 Derivation of $\kappa$

According to our pruning rule, the probability that a particular edge $w_{ji}$ is retained in the pruned set is $\kappa|w_{ji}|$. Noting the fact that all edges in the matrix $W \in \mathbb{R}^{m\times n}$ are sampled independently, the expected number of edges in the pruned matrix $W^{sparse}$ is simply the sum of probabilities that each individual edge of $W$ is retained, i.e.,

$$\mathbb{E}\left[||W^{sparse}||_0\right] = \sum_{i=1}^{m}\sum_{j=1}^{n} \kappa|w_{ji}|$$

Assuming we wish $W^{sparse}$ to have a sparsity of $s$, the number of edges in the pruned matrix needs to be $(1-s)mn$, thus giving us the equality

$$(1-s)mn = \sum_{i=1}^{m}\sum_{j=1}^{n} \kappa|w_{ji}| = \kappa \sum_{i=1}^{m}\sum_{j=1}^{n} |w_{ji}| = \kappa||W||_{L^1} \implies \kappa = \frac{(1-s)mn}{||W||_{L^1}}$$

When the matrix $W$ represents a recurrent circuit of $N$ neurons, $m = n = N \implies \kappa = \frac{(1-s)N^2}{||W||_{L^1}}$.

## 4.2 Re-normalization of sampling probabilities

Since the initial probability estimates $\kappa|w_{ij}|$ may result in values greater than 1, applying them directly could lead to an overestimation for the probability of retaining certain elements, potentially causing the final sparsity level to deviate from the target. We therefore take a renormalization step to adjust the probabilities so that they sum appropriately, enabling the pruning rule to meet the desired sparsity while maintaining consistency with probabilistic interpretation.

**Algorithm 3:** Probability Re-normalization

**Input:** arr: Array of all probabilities computed as $\kappa|w_{ij}|$

**Output:** arr: Adjusted array with re-normalized probabilities

$r \leftarrow \sum_{i=1}^{n} arr[i] - 1$ if $arr[i] > 1$ else 0 ; `// Calculate initial total residue`

**while** $r > 0$ **do**

  $counts \leftarrow \sum_{i=1}^{n} 1$ if $arr[i] < 1$ else 0 ; `// Count number of probabilities less than 1`

  $\delta \leftarrow r/counts$ ; `// Estimate delta to be added per probability < 1`

  **for** $i \leftarrow 1$ **to** $n$ **do**

    $arr[i] \leftarrow arr[i] + \delta$ if $arr[i] < 1$ else 1 ; `// Update array with delta added`

  **end**

  $r \leftarrow \sum_{i=1}^{n} arr[i] - 1$ if $arr[i] > 1$ else 0 ; `// Calculate new total residue`

**end**

**return** $arr$ ; `// Return re-normalized probability array`

## 4.3 Adjusting sparsity values

When targeting a specific sparsity level $s$ for a matrix (or block of weights), we may find that the matrix $W$ already contains a certain number of zero entries, denoted by $z_0$. If these existing zeros are not accounted for, applying the desired sparsity $s$ directly may result in a final sparsity that exceeds the intended level. Therefore, we adjust $s$ to a new value $s'$, which takes into account the current sparsity of $W$ as follows:

$$s' = \frac{s \cdot N_{total} - z_0}{N_{total} - z_0}$$

# 5 Expected overlaps with MST across sampling methods

## 5.1 Lower bounding expected overlap: Top-prob pruned network vs dense MST

Here we establish a (loose) lower bound on the expected overlap between the weights kept when sparsifying an RNN using the top-prob pruning rule and the maximum spanning tree (MST) of the original network, which from the view point of persistent homology, encapsulates all the zeroth-order topological information of a (trained) network.

Consider the square connectivity matrix with $N$ pre-synaptic and post-synaptic neurons each. For our purposes we will assume that every neuron always acts as both, a source and a target to at least one other (but not necessarily the same) neuron. The total number of weights in the connectivity matrix is then $N^2$ of which $(1-s)N^2$ will be sampled for the pruned network to have a target sparsity of $s$. The MST of such a weight matrix will have exactly $2N-1$ weights.

Following Kruskal's algorithm, the probability that the $k^{\text{th}}$ largest weight by magnitude is in the MST of the bipartite graph can be lower bounded as follows:

In a bipartite graph, the smallest cycle must have at least 4 edges. Therefore, the largest three weights ($k \leq 3$) are always in the MST.

For $k > 3$ we can lower bound the probability that the $k^{th}$ largest edge is in the MST. In particular we note that for a weight to lie in the MST, it must not form a cycle, meaning that it either joins two nodes that were both previously not connected to any other nodes in the graph, or joins at most one new node to another connected component on the graph. For the purposes of establishing a lower bound, we only look at the probability of the former.

Consequently, for $3 < k < N$

$$\mathbb{P}[k^{th} \text{ largest weight connects two isolated vertices}] \geq \frac{(N-(k-1))^2}{N^2-(k-1)}$$

Equivalently,

$$\mathbb{P}[k^{th} \text{ largest weight is in MST}] \geq \frac{(N-(k-1))^2}{N^2-(k-1)}$$

Combining the previous two statements we get the bound:

$$\mathbb{P}[k^{th} \text{ largest edge is in MST}] \geq \begin{cases} 1 & \text{for } k \leq 3 \\ \frac{(N-(k-1))^2}{N^2-(k-1)} & \text{for } 3 < k < N \\ 0 & \text{for } k \geq N \end{cases}$$

The expected number of weights that overlap with the MST is then simply bounded as

$$\begin{aligned} \mathbb{E}[N_{overlap}]_{top-prob} &\geq \sum_{k=1}^{N^2} \kappa|w_k| \cdot \mathbb{P}[k^{th} \text{ largest weight is in MST}] \\ &= \sum_{k=1}^{3} \kappa|w_k| + \sum_{k=4}^{N} \kappa|w_k| \cdot \left(\frac{(N-(k-1))^2}{N^2-(k-1)}\right) \end{aligned}$$

where $\kappa = \frac{(1-s)N^2}{||W||_{L^1}}$

## 5.2 Expected overlap with MST for random pruning

In the case of random sampling, quantifying the expected number of weights which overlap between the MST and sampled weights is equivalent to that between two arbitrarily chosen subsets, one with $2N-1$ weights (i.e., the same size as the MST) and the other with $(1-s)N^2$ weights (i.e., the same size as the sparsified connectivity matrix).

To do so we can directly use the expression for the probability mass function of a hypergeometric distribution, where the probability of a random variable $X$ having $k$ successes (random draws for which the object drawn has a specified feature) in $n$ independent draws (without replacement) from a finite population of size $M$ objects that contains exactly $K$ objects with that feature is given as

$$p_X(k) = \mathbb{P}(X = k) = \frac{\binom{K}{k} \cdot \binom{M-K}{n-k}}{\binom{M}{n}}$$

Noting that the MST is a fixed set of $2N-1$ weights for any instantiation of the connectivity matrix, the probability that it has exactly $k$ weights overlapping with the randomly sampled set of size

$(1-s)N^2$ out of a possible set of $N^2$ weights is

$$\mathbb{P}\left[N_{overlap}\right]_{random} = \frac{\binom{2N-1}{k} \cdot \binom{N^2-2N+1}{(1-s)N^2-k}}{\binom{N^2}{(1-s)N^2}}$$

The expected number of sampled weights that overlap with the MST then is simply

$$\begin{aligned}\mathbb{E}\left[N_{overlap}\right]_{random} &= \sum_{k=1}^{n} k \cdot \mathbb{P}\left[N_{overlap}\right]_{random} \\ &= \sum_{k=1}^{n} k \cdot \frac{\binom{2N-1}{k} \cdot \binom{N^2-2N+1}{(1-s)N^2-k}}{\binom{N^2}{(1-s)N^2}}\end{aligned}$$

where $n = \min(2N-1, (1-s)N^2)$.

# 6 Data pre-processing: Imaging depths

See fig S1.

# 7 Connection probabilities within and across cell populations

See table 9.

# 8 Connectivity differences across varying degrees of spatial resolution and types of error

## Connectivity differences: Familiar vs. Novel

See fig S2. In all three cases we see that the presentation of a novel image (which is a type of expectation violation) increases the inter-areal feedforward connectivity V1 L2/3 → LM L4 in the microcircuit. We see an increase in the feedback connectivity V1 ← LM as well, but the specific feedback pathway that is engaged changes with the type of error; In particular, compared to the Familiar No Change condition, during the Novel No Change condition, an increase in V1 L5 ← LM 5 dominates, while in the case of Familiar Change vs. Novel Change, V1 L2/3 ← LM 5 is the dominant form of increased inter-areal feedback. Across both sets of comparisons, we notice that the novel case leads to increased activity in V1 Vip neurons in both layers 2/3 and 5. In the case of Familiar Omission vs. Novel Omission, we note that there is an increase in the feedback V1 L2/3 ← LM L5 during the Novel Omission condition, as with the other two conditions. However in this case the projection V1 L5 ← LM L5 is lower during the Novel Omission condition, once again speaking to the specificity of feedback projections depending on the type of novelty. As before, we notice that the novel case leads to increased activity in V1 Vip neurons in both layers 2/3 and 5.

## Connectivity differences: No Change vs. Omission

See fig S3. In both cases we see that the omission condition increases the inter-areal feedforward connectivity V1 L2/3 → LM L4 as well as both forms of inter-areal V1 ← LM feedback in the

microcircuit. Broadly, both no-change vs. omission conditions seem to induce similar connectivity across the various cell-type populations and layers, indicating that the omission, i.e., type of violation of the stimulus, strongly affects the connectivity and subsequent in the microcircuit.

However, at the cell-type level, we see that the strength of interaction from Sst to Vip in L2/3 increases under conditions of novelty and omission, which initially seems at odds with the activity-based findings in (*56, 65*) that state Vip response is increased during these conditions and Sst activity is reduced. One explanation is that the inferred connectivity is correlative rather than causative, and given that this circuit motif is itself embedded in a larger network with other cell-type populations whose activities we do not have access to in this set of experiments (e.g., PV cells), their absence might skew our results at this finer level when studying inferred connectivity.

# 9 Inferring Connectivity: Control experiments

## Inferred connectivity with conventional backpropagation

See fig. S4. Comparing CelltypeRNN models trained with sign-preserving Dale's backprop (right) and those trained with conventional backprop (left) we observe that while the inferred connectivites are consistent across both training methods in the differences in inferred connectivity for the Familiar Change vs. Familiar No Change (Full presentation) condition (fig S4. A), this does not hold true in the Familiar Omission vs. Familiar Change condition (fig S4. B), albeit the difference is not much. This implies that the structure of the microcircuit itself likely drives much of macroscale dynamics that we observe, while the individual sign-constrained components modulate them at a finer scale.

## Inferred connectivity with shuffled responses

See fig. S5. To ascertain whether it is only our physical sparse connectivity constraints that are responsible for the inferred connectivities that are consistent with predictive coding, we also learn models where the overall physical structure of the microcircuit is preserved while the responses these neuronal populations learn are cyclically shuffled, i.e., L2/3 fits the responses of L4, L4 fits the

responses of L5 and L5 fits the responses of L2/3, in both areas V1 and LM. Additionally, we keep the individual cell-types consistent, i.e., neurons indexed as Pyr neurons still fit to the responses of Pyr neurons, just from a different layer, and the same holds true for Sst and Vip neurons.

Comparing CelltypeRNN models trained to fit neuronal responses that are "shuffled" (left), vs. those that are trained to fit their correct unshuffled responses (right), we observe that the inferred connectivities are consistent across both in the differences in inferred connectivity for the Familiar Change vs. Familiar No Change (Full presentation) condition (fig S5. A). However, this does not hold true in either the Familiar Change vs. Familiar No Change (Half presentation) condition (fig S5. B), or the Familiar Omission vs. Familiar Change condition fig S5. C), giving added credence to our original experimental results. In particular, we find that in the shuffled Half session experiment both feedforward and feedback connections are higher in the No Change case, therefore no longer being indicative of the differences in timescales at which they operate, while in the shuffled Familiar Omission vs. Familiar Change case, both feedback connections from LM to V1 are higher higher in the Change case vs. the Omission case, going against the hypothesis that greater expectation violation would lead to an increase in feedback activity.

# Supplementary figures

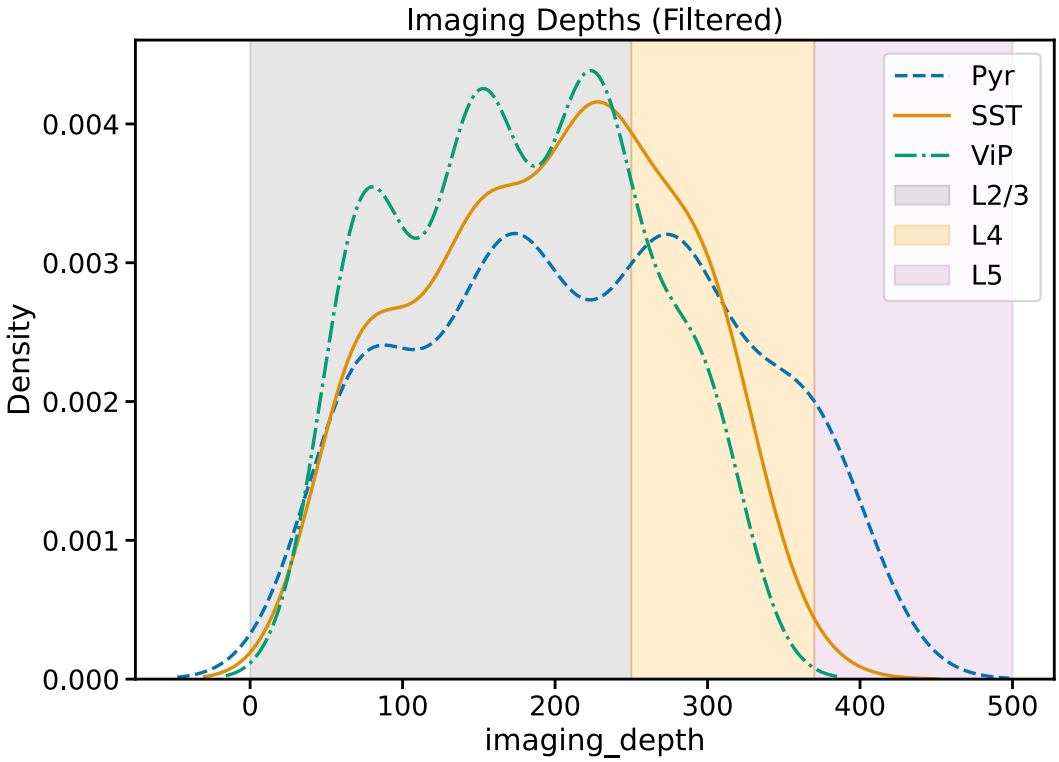


**Figure S1**: **Imaging depths.** Distribution of experiments (Y-axis) with respect to imaging depth (X-axis). Shaded grey, yellow and pink panels represent depths corresponding to cortical layers 2/3, 4, and 5 respectively.

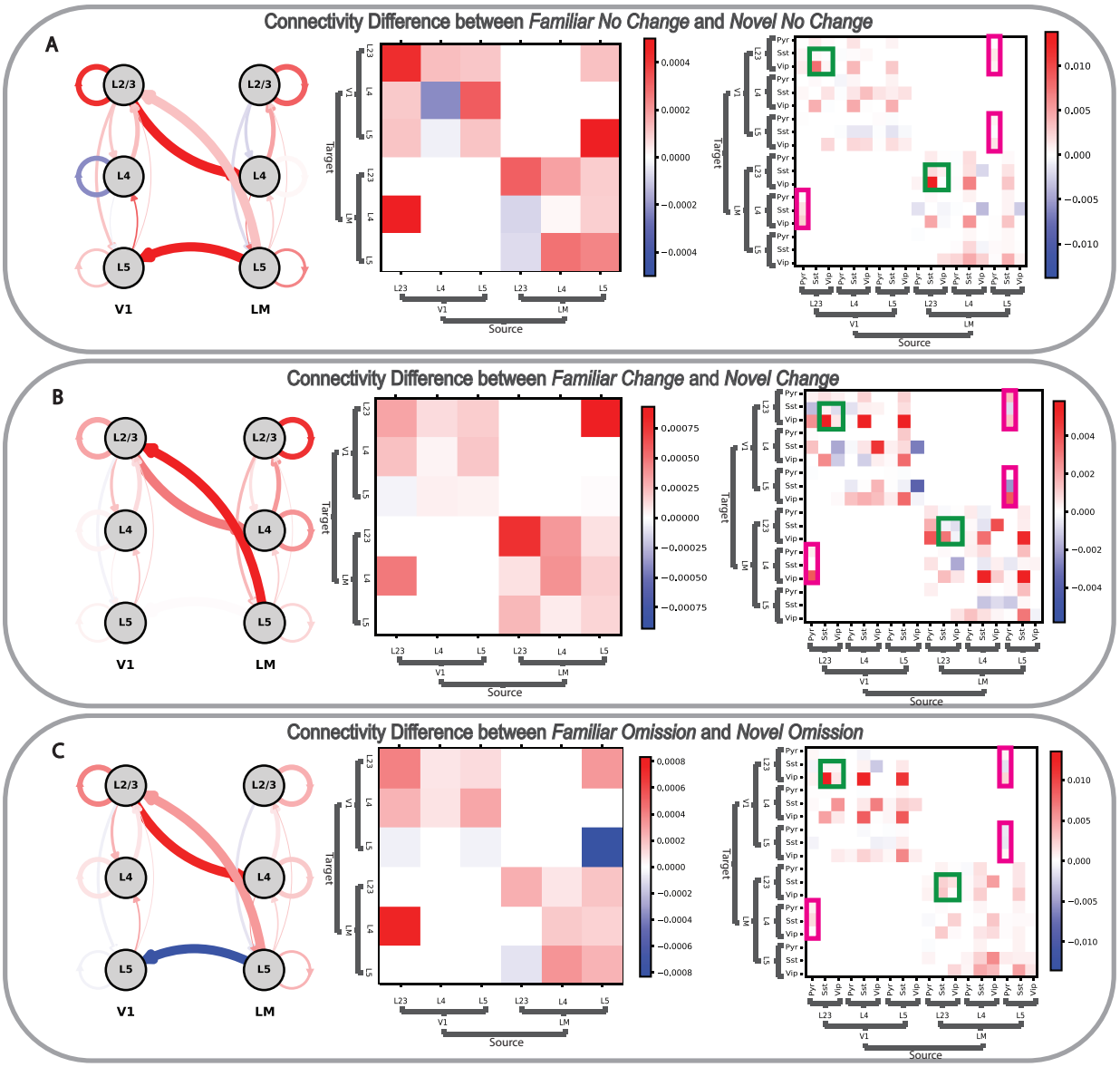


**Figure S2**: **Connectivity differences between familiar and novel conditions (Full presentations). (A)** Familiar no change vs. Novel no change. **(B)** Familiar change vs. Novel change. **(C)** Familiar omission vs. Novel omission. All plots are from the full presentations condition. All differences are computed as `Second condition - First condition`; Blue implies higher weights in the first condition, while red indicates higher weights in the second. In all cases, the left and middle plots are a graphical representation of the weights averaged across layers, while the rightmost plot averages weights by cell-type within each layer. Magenta boxes highlight the feedforward and feedback connections, i.e., those originating at V1 L2/3 and LM L5 respectively. Green boxes highlight Sst-ViP interactions in L2/3 of V1 and LM.

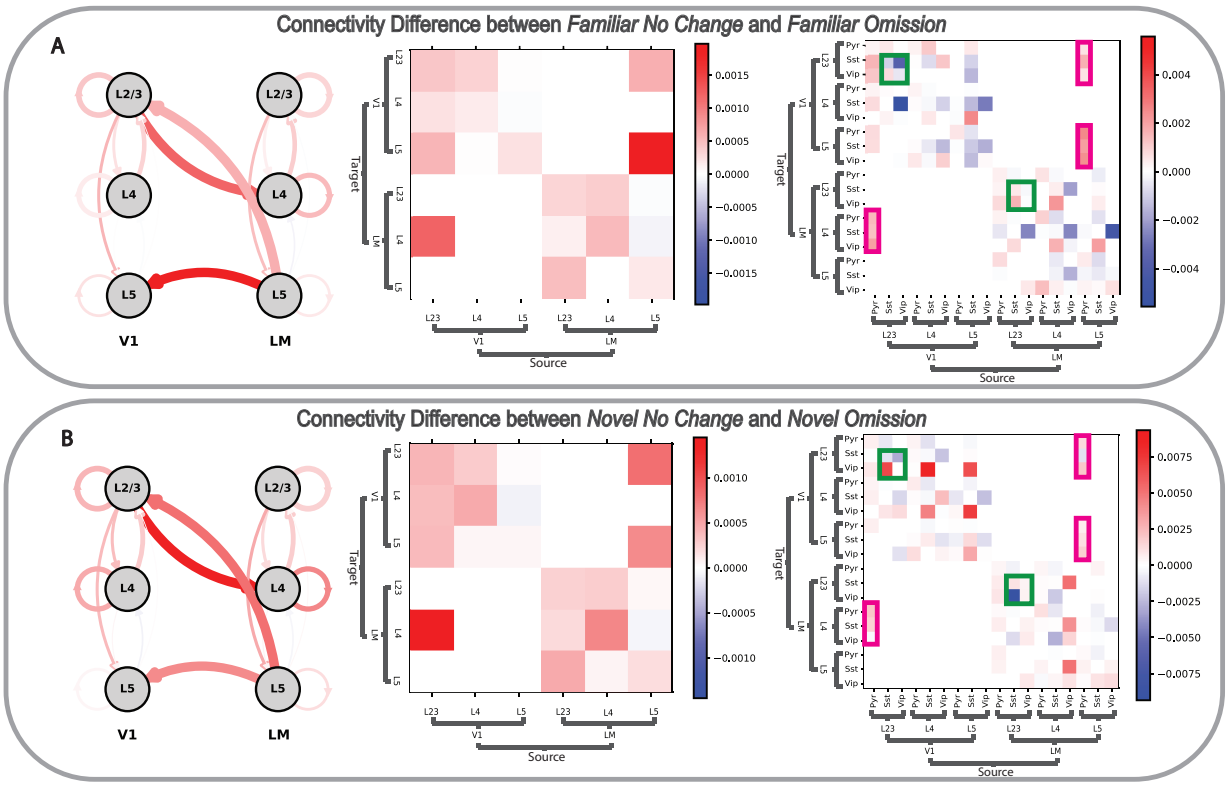


**Figure S3**: **Connectivity differences between no change and omission conditions (Full presentations).** **(A)** Familiar No Change vs. Familiar Omission. **(B)** Novel No Change vs. Novel Omission. Both plots are from the full presentations condition. All plots are from the full presentations condition. Differences are computed as `Second condition - First condition`; Blue implies higher weights in the first condition, while red indicates higher weights in the second. In both cases, the left and middle plots are a graphical representation of the weights averaged across layers, while the rightmost plot averages weights by cell-type within each layer. Magenta boxes highlight the feedforward and feedback connections, i.e., those originating at V1 L2/3 and LM L5 respectively. Green boxes highlight Sst-ViP interactions in L2/3 of V1 and LM.

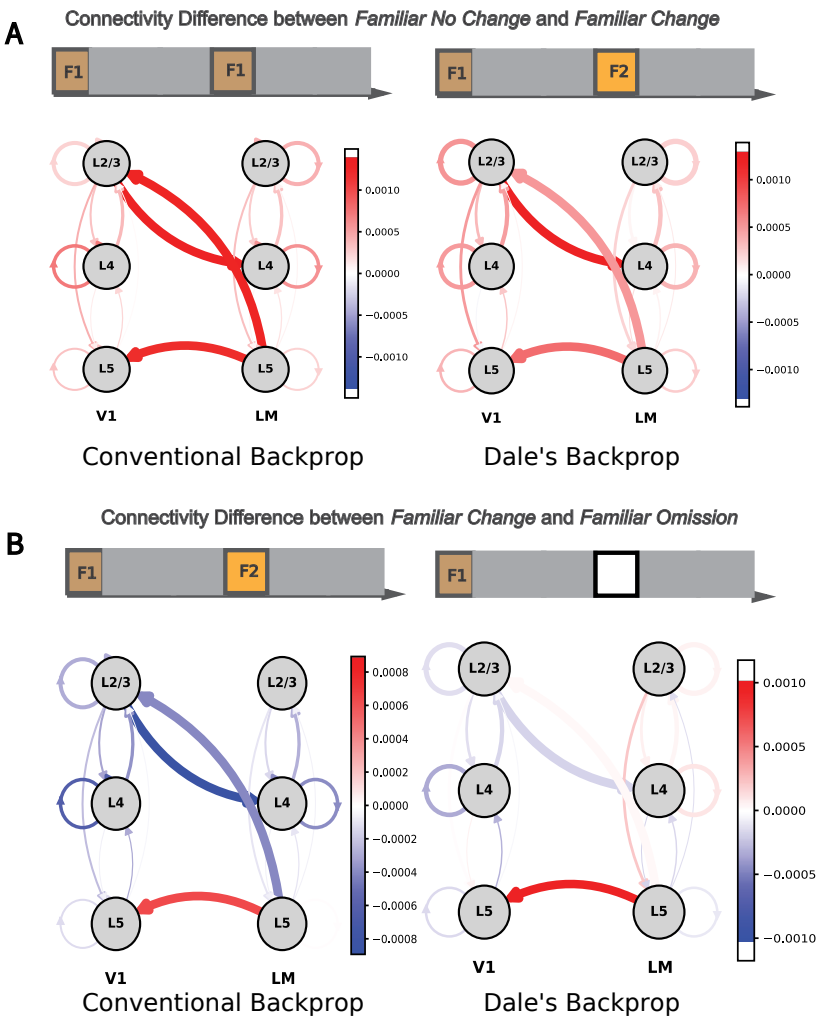


**Figure S4**: **Connectivity differences compared with vanilla backprop controls.** **(A)** Familiar No Change vs. Familiar Change (Full presentation). **(B)** Familiar No Change vs. Familiar Omission (Full presentation). In all cases, figures on the left represent inferred connectivity in the shuffled case while the figure on the right represents inferred connectivity in the unshuffled case.

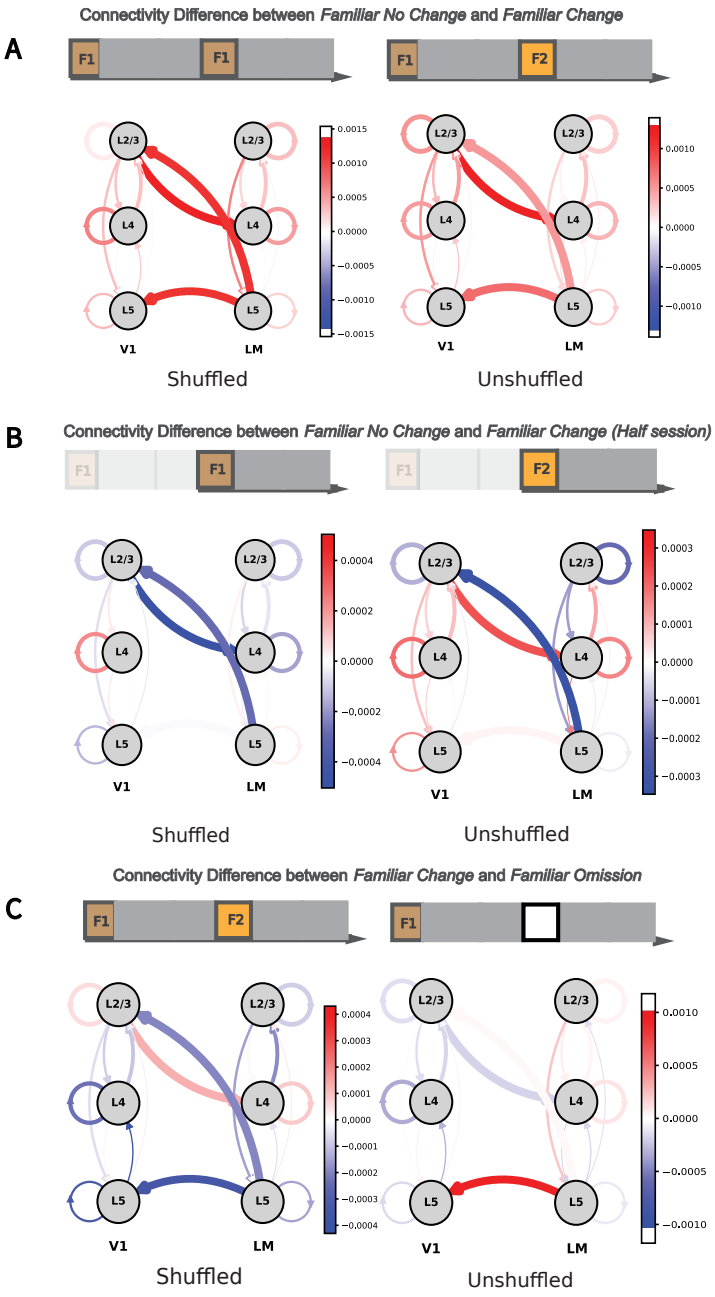


**Figure S5**: **Connectivity differences with shuffled controls.** **(A)** Familiar No Change vs. Familiar Change (Full presentation). **(B)** Familiar No Change vs. Familiar Change (Half presentation). **(C)** Familiar No Change vs. Familiar Omission (Full presentation). In all cases, figures on the left represent inferred connectivity in the shuffled case while the figure on the right represents inferred connectivity in the unshuffled case.

# Supplementary tables

| Connection Probabilities | L2/3 Pyr | L2/3 SST | L2/3 VIP | L4 Pyr | L4 SST | L4 VIP | L5 Pyr | L5 SST | L5 VIP |
|---|---|---|---|---|---|---|---|---|---|
| L2/3 Pyr | 0.06 | 0.23 | 0.05 | 0.07 | 0.33 | 0.00 | 0.00 | 0.15 | 0.00 |
| L2/3 SST | 0.3 | 0.05 | 0.14 | 0.04 | 0.05 | 0.25 | 0.00 | 0.05 | 0.00 |
| L2/3 VIP | 0.16 | 0.30 | 0.01 | 0.02 | 0.21 | 0.00 | 0.00 | 0.13 | 0.00 |
| L4 Pyr | 0.05 | 0.04 | 0.00 | 0.10 | 0.22 | 0.00 | 0.00 | 0.19 | 0.00 |
| L4 SST | 0.17 | 0.00 | 0.14 | 0.04 | 0.01 | 0.14 | 0.08 | 0.04 | 0.2 |
| L4 VIP | 0.00 | 0.13 | 0.05 | 0.00 | 0.22 | 0.03 | 0.02 | 0.14 | 0.00 |
| L5 Pyr | 0.08 | 0.00 | 0.00 | 0.00 | 0.08 | 0.00 | 0.04 | 0.15 | 0.00 |
| L5 SST | 0.04 | 0.00 | 0.00 | 0.03 | 0.04 | 0.11 | 0.10 | 0.03 | 0.05 |
| L5 VIP | 0.00 | 0.00 | 0.04 | 0.11 | 0.07 | 0.04 | 0.02 | 0.1 | 0.03 |

**Table S1**: Connection probabilities amongst different cell types and populations as per Campagnola et al. (*66*) with **Columns** = Pre-synaptic population **(source)**, **Rows** = Post-synaptic population **(target)**.

**REFERENCES AND NOTES**


1. Y. Cohen, T. A. Engel, C. Langdon, G. W. Lindsay, T. Ott, M. A. K. Peters, J. M. Shine, V. Breton-Provencher, S. Ramaswamy, Recent advances at the interface of neuroscience and artificial neural networks. *J. Neurosci.* **42**, 8514–8523 (2022).

2. A. Saxe, S. Nelli, C. Summerfield, If deep learning is the answer, what is the question? *Nat. Rev. Neurosci.* **22**, 55–67 (2021).

3. B. A. Richards, T. P. Lillicrap, P. Beaudoin, Y. Bengio, R. Bogacz, A. Christensen, C. Clopath, R. P. Costa, A. de Berker, S. Ganguli, C. J. Gillon, D. Hafner, A. Kepecs, N. Kriegeskorte, P. Latham, G. W. Lindsay, K. D. Miller, R. Naud, C. C. Pack, P. Poirazi, P. Roelfsema, J. Sacramento, A. Saxe, B. Scellier, A. C. Schapiro, W. Senn, G. Wayne, D. Yamins, F. Zenke, J. Zylberberg, D. Therien, K. P. Kording, A deep learning framework for neuroscience. *Nat. Neurosci.* **22**, 1761–1770 (2019).

4. T. C. Kietzmann, P. McClure, N. Kriegeskorte, Deep neural networks in computational neuroscience. *Oxford Res. Encycl. Neurosci.* 10.1093/acrefore/9780190264086.013.46 (2019).

5. D. L. Yamins, J. J. DiCarlo, Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **19**, 356–365 (2016).

6. O. Barak, Recurrent neural networks as versatile tools of neuroscience research. *Curr. Opin. Neurobiol.* **46**, 1–6 (2017).

7. G. R. Yang, X.-J. Wang, Artificial neural networks for neuroscientists: A primer. *Neuron* **107**, 1048–1070 (2020).

8. M. T. Kaufman, M. M. Churchland, S. I. Ryu, K. V. Shenoy, Cortical activity in the null space: Permitting preparation without movement. *Nat. Neurosci.* **17**, 440–448 (2014).

9. M. G. Perich, J. A. Gallego, L. E. Miller, A neural population mechanism for rapid learning. *Neuron* **100**, 964–976.e7 (2018).

10. J. D. Semedo, A. Zandvakili, C. K. Machens, M. Y. Byron, A. Kohn, Cortical areas interact through a communication subspace. *Neuron* **102**, 249–259.e4 (2019).

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

12. L. Kozachkov, M. Ennis, J.-J. Slotine, RNNs of RNNs: Recursive construction of stable assemblies of recurrent neural networks. *Adv. Neural Inf. Process. Syst.* **35**, 30512–30527 (2022).

13. D. Sussillo, L. F. Abbott, Generating coherent patterns of activity from chaotic neural networks. *Neuron* **63**, 544–557 (2009).

14. B. DePasquale, C. J. Cueva, K. Rajan, G. S. Escola, L. Abbott, full-FORCE: A target-based method for training recurrent networks. *PLOS ONE* **13**, e0191527 (2018).

15. D. Sussillo, M. M. Churchland, M. T. Kaufman, K. V. Shenoy, A neural network that finds a naturalistic solution for the production of muscle activity. *Nat. Neurosci.* **18**, 1025–1033 (2015).

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

17. C. Pandarinath, D. J. O'Shea, J. Collins, R. Jozefowicz, S. D. Stavisky, J. C. Kao, E. M. Trautmann, M. T. Kaufman, S. I. Ryu, L. R. Hochberg, J. M. Henderson, K. V. Shenoy, L. F. Abbott, D. Sussillo, Inferring single-trial neural population dynamics using sequential auto-encoders. *Nat. Methods* **15**, 805–815 (2018).

18. A. A. Russo, R. Khajeh, S. R. Bittner, S. M. Perkins, J. P. Cunningham, L. F. Abbott, M. M. Churchland, Neural trajectories in the supplementary motor area and motor cortex exhibit distinct geometries, compatible with different classes of computation. *Neuron* **107**, 745–758.e6 (2020).

19. K. Rajan, C. D. Harvey, D. W. Tank, Recurrent network models of sequence generation and memory. *Neuron* **90**, 128–142 (2016).

20. M. G. Perich, C. Arlt, S. Soares, M. E. Young, C. P. Mosher, J. Minxha, E. Carter, U. Rutishauser, P. H. Rudebeck, C. D. Harvey, K. Rajan, Inferring brain-wide interactions using data-constrained recurrent neural network models. bioRxiv 423348 [Preprint] (2020). https://doi.org/10.1101/2020.12.18.423348.

21. N. Maheswaranathan, A. Williams, M. Golub, S. Ganguli, D. Sussillo, Reverse engineering recurrent networks for sentiment classification reveals line attractor dynamics. *Adv. Neural Inf. Process. Syst.* **32** (2019).

22. D. Sussillo, O. Barak, Opening the black box: Low-dimensional dynamics in high-dimensional recurrent neural networks. *Neural Comput.* **25**, 626–649 (2013).

23. D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, J. J. Di Carlo, Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 8619–8624 (2014).

24. A. J. Kell, D. L. Yamins, E. N. Shook, S. V. Norman-Haignere, J. H. McDermott, A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron* **98**, 630–644.e16 (2018).

25. J. Kubilius, M. Schrimpf, K. Kar, R. Rajalingham, H. Hong, N. J. Majaj, E. B. Issa, P. Bashivan, J. Prescott-Roy, K. Schmidt, A. Nayebi, D. Bear, D. L. K. Yamins, J. J. Di Carlo, Brain-like object recognition with high-performing shallow recurrent ANNs. *Adv. Neural Inf. Process. Syst.* **32** (2019).

26. M. Schrimpf, J. Kubilius, H. Hong, N. J. Majaj, R. Rajalingham, E. B. Issa, K. Kar, P. Bashivan, J. Prescott-Roy, K. Schmidt, D. L. K. Yamins, James J. Di Carlo, Brain-score: Which artificial neural network for object recognition is most brain-like? bioRxiv 407007 [Preprint] (2018). https://doi.org/10.1101/407007.

27. J. A. Michaels, S. Schaffelhofer, A. Agudelo-Toro, H. Scherberger, A modular neural network model of grasp movement generation. bioRxiv 742189 [Preprint] (2019). https://doi.org/10.1101/742189.

28. A. Nayebi, D. Bear, J. Kubilius, K. Kar, S. Ganguli, D. Sussillo, James J. Di Carlo, D. L. K. Yamins, Task-driven convolutional recurrent models of the visual system. *Adv. Neural Inf. Process. Syst.* **31** (2018).

29. G. W. Lindsay, Convolutional neural networks as a model of the visual system: Past, present, and future. *J. Cogn. Neurosci.* **33**, 2017–2031 (2021).

30. D. Hassabis, D. Kumaran, C. Summerfield, M. Botvinick, Neuroscience-inspired artificial intelligence. *Neuron* **95**, 245–258 (2017).

31. R. Schaeffer, M. Khona, I. Fiete, No free lunch from deep learning in neuroscience: A case study through models of the entorhinal-hippocampal circuit. *Adv. Neural Inf. Process. Syst.* **35**, 16052–16067 (2022).

32. A. H. Marblestone, G. Wayne, K. P. Kording, Toward an integration of deep learning and neuroscience. *Front. Comput. Neurosci.* **10**, 94 (2016).

33. J. C. Eccles, From electrical to chemical transmission in the central nervous system: The closing address of the Sir Henry Dale centennial symposium, Cambridge, 19 September 1975. *Notes Rec. R. Soc. Lond.* **30**, 219–230 (1976).

34. H. Eavani, T. D. Satterthwaite, R. Filipovych, R. E. Gur, R. C. Gur, C. Davatzikos, Identifying sparse connectivity patterns in the brain using resting-state fMRI. *Neuroimage* **105**, 286–299 (2015).

35. M. Kaiser, Connectomes: From a sparsity of networks to large-scale databases. *Front. Neuroinform.* **17**, 1170337 (2023).

36. J. K. Lappalainen, F. D. Tschopp, S. Prakhya, M. M. Gill, A. Nern, K. Shinomiya, S.-Y. Takemura, E. Gruntman, J. H. Macke, S. C. Turaga, Connectome-constrained networks predict neural activity across the fly visual system. *Nature* **634**, 1132–1140 (2024).

37. G. Giacopelli, D. Tegolo, E. Spera, M. Migliore, On the structural connectivity of large-scale models of brain networks at cellular level. *Sci. Rep.* **11**, 4345 (2021).

38. J. Cornford, D. Kalajdzievski, M. Leite, A. Lamarquette, D. M. Kullmann, B. Richards, Learning to live with Dale's principle: ANNs with separate excitatory and inhibitory units. *Int. Conf. Learn. Represent.* 1–27 (2021).

39. J. Frankle, M. Carbin, The lottery ticket hypothesis: Finding sparse, trainable neural networks. *Int. Conf. Learn. Represent.* 1–42 (2019).

40. H. Tanaka, D. Kunin, D. L. Yamins, S. Ganguli, Pruning neural networks without any data by iteratively conserving synaptic flow. *Adv. Neural Inf. Process. Syst.* **33**, 6377–6389 (2020).

41. N. Lee, T. Ajanthan, P. H. Torr, SNIP: Single-shot network pruning based on connection sensitivity. arXiv:1810.02340 [cs.CV] (2018).

42. C. Wang, G. Zhang, R. Grosse, Picking winning tickets before training by preserving gradient flow. arXiv:2002.07376 [cs.LG] (2020).

43. S. Han, J. Pool, J. Tran, W. Dally, Learning both weights and connections for efficient neural network. *Adv. Neural Inf. Process. Syst.* **28** (2015).

44. T. Miconi, Biologically plausible learning in recurrent neural networks reproduces neural dynamics observed during cognitive tasks. *eLife* **6**, e20899 (2017).

45. S. Minni, L. Ji-An, T. Moskovitz, G. Lindsay, K. Miller, M. Dipoppa, G. R. Yang, Understanding the functional and structural differences across excitatory and inhibitory neurons. bioRxiv 680439 [Preprint] (2019). https://doi.org/10.1101/680439.

46. A. Ingrosso, L. Abbott, Training dynamically balanced excitatory-inhibitory networks. *PLOS ONE* **14**, e0220547 (2019).

47. W. Nicola, C. Clopath, Supervised learning in spiking neural networks with FORCE training. *Nat. Commun.* **8**, 2208 (2017).

48. Y. LeCun, J. Denker, S. Solla, Optimal brain damage. *Adv. Neural Inf. Process. Syst.* **2** (1989).

49. E. Moore, R. Chaudhuri, Using noise to probe recurrent neural network structure and prune synapses. *Adv. Neural Inf. Process. Syst.* **33**, 14046–14057 (2020).

50. H. F. Song, G. R. Yang, X.-J. Wang, Training excitatory-inhibitory recurrent neural networks for cognitive tasks: A simple and flexible framework. *PLOS Comput. Biol.* **12**, e1004792 (2016).

51. D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986).

52. L. Zhang, X. Wang, R. Cueto, C. Effi, Y. Zhang, H. Tan, X. Qin, Y. Ji, X. Yang, H. Wang, Biochemical basis and metabolic interplay of redox regulation. *Redox Biol.* **26**, 101284 (2019).

53. C. Tetzlaff, C. Kolodziejski, M. Timme, F. Wörgötter, Synaptic scaling in combination with many generic plasticity mechanisms stabilizes circuit connectivity. *Front. Comput. Neurosci.* **5**, 47 (2011).

54. P. R. Huttenlocher, Synaptic density in human frontal cortex-developmental changes and effects of aging. *Brain Res.* **163**, 195–205 (1979).

55. E. Bullmore, O. Sporns, The economy of brain network organization. *Nat. Rev. Neurosci.* **13**, 336–349 (2012).

56. M. Garrett, S. Manavi, K. Roll, D. R. Ollerenshaw, P. A. Groblewski, N. D. Ponvert, J. T. Kiggins, L. Casal, K. Mace, A. Williford, A. Leon, X. Jia, P. Ledochowitsch, M. A. Buice, W. Wakeman, S. Mihalas, S. R. Olsen, Experience shapes activity dynamics and stimulus coding of VIP inhibitory cells. *eLife* **9**, e50340 (2020).

57. R. P. Rao, D. H. Ballard, Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2**, 79–87 (1999).

58. J. C. Ye, *Geometry of Deep Learning* (Springer, 2022).

59. P. Li, J. Cornford, A. Ghosh, B. Richards, Learning better with Dale's law: A spectral perspective. *Adv. Neural Inf. Process. Syst.* **36** (2024).

60. A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. De Vito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in pytorch. *Adv. Neural Inf. Process. Syst.* **30** (2017).

61. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. De Vito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32** (2019).

62. B. Rieck, M. Togninalli, C. Bock, M. Moor, M. Horn, T. Gumbsch, K. Borgwardt, Neural persistence: A complexity measure for deep neural networks using algebraic topology. arXiv:1812.09764 [cs.LG] (2018).

63. K. Friston, A theory of cortical responses. *Philos. Trans. R. Soc. B Biol. Sci.* **360**, 815–836 (2005).

64. Visual Behavior—2p. https://portal.brain-map.org/circuits-behavior/visual-behavior-2p [accessed 20 November 2024].

65. M. Garrett, P. Groblewski, A. Piet, D. Ollerenshaw, F. Najafi, I. Yavorska, A. Amster, C. Bennett, M. Buice, S. Caldejon, L. Casal, F. D'Orazi, S. Daniel, S. E. J. de Vries, D. Kapner, J. Kiggins, J. Lecoq, P. Ledochowitsch, S. Manavi, N. Mei, C. B. Morrison, S. Naylor, N. Orlova, J. Perkins, N. Ponvert, C. Roll, S. Seid, D. Williams, A. Williford, R. Ahmed, D. Amine, Y. Billeh, C. Bowman, N. Cain, A. Cho, T. Dawe, M. Departee, M. Desoto, D. Feng, S. Gale, E. Gelfand, N. Gradis, C. Grasso, N. Hancock, B. Hu, R. Hytnen, X. Jia, T. Johnson, I. Kato, S. Kivikas, L. Kuan, Q. L'Heureux, S. Lambert, A. Leon, E. Liang, F. Long, K. Mace, I. M. de Abril, C. Mochizuki, C. Nayan, K. North, L. Ng, G. K. Ocker, M. Oliver, P. Rhoads, K. Ronellenfitch, K. Schelonka, J. Sevigny, D. Sullivan, B. Sutton, J. Swapp, T. K. Nguyen, X. Waughman, J. Wilkes, M. Wang, C. Farrell, W. Wakeman, H. Zeng, J. Phillips, S. Mihalas, A. Arkhipov, C. Koch, S. R. Olsen, Stimulus novelty uncovers coding diversity in survey of visual cortex. bioRxiv 528085 [Preprint] (2023). https://doi.org/10.1101/2023.02.14.528085.

66. L. Campagnola, S. C. Seeman, T. Chartrand, L. Kim, A. Hoggarth, C. Gamlin, S. Ito, J. Trinh, P. Davoudian, C. Radaelli, M. H. Kim, T. Hage, T. Braun, L. Alfiler, J. Andrade, P. Bohn, R.

Dalley, A. Henry, S. Kebede, A. Mukora, D. Sandman, G. Williams, R. Larsen, C. Teeter, T. L. Daigle, K. Berry, N. Dotson, R. Enstrom, M. Gorham, M. Hupp, S. Dingman Lee, K. Ngo, P. R. Nicovich, L. Potekhina, S. Ransford, A. Gary, J. Goldy, D. McMillen, T. Pham, M. Tieu, L.'A. Siverts, M. Walker, C. Farrell, M. Schroedter, C. Slaughterbeck, C. Cobb, R. Ellenbogen, R. P. Gwinn, C. D. Keene, A. L. Ko, J. G. Ojemann, D. L. Silbergeld, D. Carey, T. Casper, K. Crichton, M. Clark, N. Dee, L. Ellingwood, J. Gloe, M. Kroll, J. Sulc, H. Tung, K. Wadhwani, K. Brouner, T. Egdorf, M. Maxwell, M. McGraw, C. A. Pom, A. Ruiz, J. Bomben, D. Feng, N. Hejazinia, S. Shi, A. Szafer, W. Wakeman, J. Phillips, A. Bernard, L. Esposito, F. D. D'Orazi, S. Sunkin, K. Smith, B. Tasic, A. Arkhipov, S. Sorensen, E. Lein, C. Koch, G. Murphy, H. Zeng, T. Jarsky, Local connectivity and synaptic dynamics in mouse and human neocortex. *Science* **375**, eabj5861 (2022).

67. A. Schulz, C. Miehl, M. J. Berry II, J. Gjorgjieva, The generation of cortical novelty responses through inhibitory plasticity. *eLife* **10**, e65309 (2021).

68. A. M. Bastos, W. M. Usrey, R. A. Adams, G. R. Mangun, P. Fries, K. J. Friston, Canonical microcircuits for predictive coding. *Neuron* **76**, 695–711 (2012).

69. C. A. Bosman, J. M. Schoffelen, N. Brunet, R. Oostenveld, A. M. Bastos, T. Womelsdorf, B. Rubehn, T. Stieglitz, P. de Weerd, P. Fries, Attentional stimulus selection through selective synchronization between monkey visual areas. *Neuron* **75**, 875–888 (2012).

70. J. D. Semedo, A. I. Jasper, A. Zandvakili, A. Krishna, A. Aschner, C. K. Machens, A. Kohn, B. M. Yu, Feedforward and feedback interactions between visual cortical areas use different population activity patterns. *Nat. Commun.* **13**, 1099 (2022).

71. J.-Y. Moon, K. Müsch, C. E. Schroeder, T. A. Valiante, C. J. Honey, Inter-regional delays fluctuate in the human cerebral cortex. *eLife* **13**, RP92459 (2024).

72. A. Balwani, S. Cho, H. Choi, Exploring the architectural biases of the cortical microcircuit. *Neural Comput.* **37**, 1551–1599 (2025).

73. F. Najafi, S. Russo, J. Lecoq, Unexpected events trigger task-independent signaling in VIP and excitatory neurons of mouse visual cortex. *iScience* **28**, 111728 (2025).

74. L. Hertäg, H. Sprekeler, Learning prediction error neurons in a canonical interneuron circuit. *eLife* **9**, e57541 (2020).

75. J. Haarsma, P. C. Fletcher, J. D. Griffin, H. J. Taverne, H. Ziauddeen, T. J. Spencer, C. Miller, T. Katthagen, I. Goodyer, K. M. J. Diederen, G. K. Murray, Precision weighting of cortical unsigned prediction error signals benefits learning, is mediated by dopamine, and is impaired in psychosis. *Mol. Psychiatry* **26**, 5320–5333 (2021).

76. C. K. Starkweather, N. Uchida, Dopamine signals as temporal difference errors: Recent advances. *Curr. Opin. Neurobiol.* **67**, 95–105 (2021).

77. L. Hertäg, C. Clopath, Prediction-error neurons in circuits with multiple neuron types: Formation, refinement, and functional implications. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2115699119 (2022).

78. B. A. Richards, T. P. Lillicrap, Dendritic solutions to the credit assignment problem. *Curr. Opin. Neurobiol.* **54**, 28–36 (2019).

79. J. A. Westerberg, Y. S. Xiong, E. Sennesh, H. Nejat, D. Ricci, S. Durand, B. Hardcastle, H. Cabasco, H. Belski, A. Bawany, R. Gillis, H. Loeffler, C. R. Peene, W. Han, K. Nguyen, V. Ha, T. Johnson, C. Grasso, A. Young, J. Swapp, B. Ouellette, S. Caldejon, A. Williford, P. A. Groblewski, S. R. Olsen, C. Kiselycznyk, C. Koch, J. A. Lecoq, A. Maier, A. M. Bastos, Stimulus history, not expectation, drives sensory prediction errors in mammalian cortex. *bioRxiv* 616378 [Preprint] (2024). https://doi.org/10.1101/2024.10.02.616378.

80. S. Furutachi, A. D. Franklin, A. M. Aldea, T. Mrsic-Flogel, S. B. Hofer, Cooperative thalamocortical circuit mechanism for sensory prediction errors. *Adv. Neural Inf. Process. Syst.* **633**, 398–406 (2024).

81. J. L. Elman, Finding structure in time. *Cognit. Sci.* **14**, 179–211 (1990).

82. S. P. Boyd, L. Vandenberghe, *Convex Optimization* (Cambridge Univ. Press, 2004).

83. D. P. Bertsekas, *Nonlinear Programming* (Athena Scientific, ed. 3, 2016).

84. M. C. Mozer, P. Smolensky, Skeletonization: A technique for trimming the fat from a network via relevance assessment. *Adv. Neural Inf. Process. Syst.* **1**, 107–115 (1988).

85. S. Hanson, L. Pratt, Comparing biases for minimal network construction with back-propagation. *Adv. Neural Inf. Process. Syst.* **1**, 177–185 (1988).

86. D. A. Spielman, N. Srivastava, "Graph sparsification by effective resistances" in *Proceedings of the Fortieth Annual ACM Symposium on Theory of computing* (ACM, 2008), pp. 563–568.

87. D. A. Spielman, S.-H. Teng, Spectral sparsification of graphs. *SIAM J. Comput.* **40**, 981–1025 (2011).

88. J. Batson, D. A. Spielman, N. Srivastava, S.-H. Teng, Spectral sparsification of graphs: Theory and algorithms. *Commun. ACM* **56**, 87–94 (2013).

89. H. Doraiswamy, J. Tierny, P. J. Silva, L. G. Nonato, C. Silva, Topomap: A 0-dimensional homology preserving projection of high-dimensional data. *IEEE Trans. Vis. Comput. Graph.* **27**, 561–571 (2020).

90. T. Lacombe, Y. Ike, M. Carriere, F. Chazal, M. Glisse, Y. Umeda, Topological uncertainty: Monitoring trained neural networks through persistence of activation graphs. arXiv:2105.04404 [stat.ML] (2021).

91. A. Balwani, J. Krzyston, "Zeroth-order topological insights into iterative magnitude pruning" in *Topological, Algebraic and Geometric Learning Workshops 2022* (PMLR, 2022), pp. 6–16.

92. P. A. Groblewski, D. R. Ollerenshaw, J. T. Kiggins, M. E. Garrett, C. Mochizuki, L. Casal, S. Cross, K. Mace, J. Swapp, S. Manavi, D. Williams, S. Mihalas, S. R. Olsen, Characterization of learning, motivation, and visual perception in five transgenic mouse lines expressing GCaMP in distinct cell populations. *Front. Behav. Neurosci.* **14**, 104 (2020).

93. R. J. Douglas, K. A. Martin, D. Whitteridge, A canonical microcircuit for neocortex. *Neural Comput.* **1**, 480–488 (1989).

94. V. B. Mountcastle, The columnar organization of the neocortex. *Brain* **120**, 701–722 (1997).

95. R. Schneider, *Convex Bodies: The Brunn–Minkowski Theory*, vol. 151 (Cambridge Univ. Press, 2013).

96. H. Kim, G. Papamakarios, A. Mnih, "The Lipschitz constant of self-attention" in *International Conference on Machine Learning* (PMLR, 2021), pp. 5562–5571.

97. H. Federer, *Geometric Measure Theory* (Springer, 2014).