Mechanistic Interpretability for Recurrent Computation: Time-Resolved Circuit Discovery in RNNs

Aishwarya Balwani [α]

Abstract

Despite being one of computational neuroscience's most prominent modelling tools, recurrent neural networks (RNNs) have shown limited utility for revealing explicit structure-function relationships in neuronal circuits. This shortcoming reflects the fact that recurrent computations are distributed across neurons, timesteps, and internal states, as a consequence of which, static summaries of weights or average activity often fail to reveal the transient causal interactions through which behavior may be implemented. In this work, we present a circuit discovery framework that adapts causal intervention techniques from mechanistic interpretability of large language models to the recurrent, multi-step computation in RNNs. By combining windowed ablations with Jacobian-based linearization of the hidden state trajectories, we estimate effective connectivity across the RNN as it evolves, thus revealing how task-relevant computations are implemented through dynamically coordinated subcircuits. Across synthetic tasks with known mechanisms, we recover operative circuits with high precision while demonstrating robustness advantages over correlation-based selection methods. On applying our methods to anatomically-constrained RNNs trained on the Allen Institute's Visual Behavior dataset, we recover VIP neuronal involvement in unexpected-stimulus processing and reveal temporally specific causal contributions invisible to static analyses. Our experiments further suggest that intrinsic timing of VIP neurons drives prediction error formation in the network, consequently bridging the timing-code and predictive-processing interpretations of their function.

[α]Department of Developmental Neurobiology, St. Jude Children's Research Hospital. Correspondence to: Aishwarya Balwani <aishwarya.balwani@stjude.org>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

1. Introduction

The advent of the deep learning (DL) revolution (Richards et al., 2019; Saxe et al., 2021) has seen recurrent neural networks (RNNs) cement their place as modelling tools in computational neuroscience (Mante et al., 2013; Barak, 2017; Yang & Wang, 2020). With numerous methods available for training these networks so that they can successfully capture the latent dynamics found in neuroscience data (Sussillo & Abbott, 2009; Song et al., 2016; DePasquale et al., 2018; Li et al., 2023; Balwani et al., 2025b), RNNs are widely used to study various aspects of brain function and intelligence [1] across different the behavioural and representational levels (Sussillo et al., 2015; Rajan et al., 2016; Pandarinath et al., 2018; Russo et al., 2020).

1.1. Background and Related Work

Although RNNs offer unique advantages as in silico experimental platforms given that their i) task-solving strategies are learnt via optimization (Cao & Yamins, 2021), and ii) full mechanistic substrate is, in principle, accessible (Cao & Yamins, 2024), our ability to interpret these models remains limited in substantive ways.

Existing RNN interpretability approaches operate largely at the level of representational geometry (Recanatesi et al., 2021; Balwani et al., 2025a), latent dynamics (Sussillo & Barak, 2013; Mastrogiuseppe & Ostojic, 2018; Maheswaranathan et al., 2019; Driscoll et al., 2024) and correlational weight structure (Perich et al., 2020). Recent work has also begun to bridge latent representations back to individual neurons through attribution methods with identifiability guarantees (Schneider et al., 2025), yet a critical gap remains: current methods fail to provide time-resolved, causal structure-function maps linking concrete circuit ele-

[1]Although transformer-based models now dominate the machine learning landscape, RNNs remain the model class of choice for relating cortical structure and neural computation. As nonlinear dynamical systems reflecting the predominantly recurrent architecture of cortex itself (Douglas et al., 1995; Douglas & Martin, 2004; 2007), representations from RNNs can be more plausibly compared to experimentally recorded population activity. Transformers and other feedforward models afford neither this dynamical correspondence nor a clean mapping of their components onto individual biological neurons.

ments to transient computation. While they can reveal which latent units correlate with task-relevant variables, they fall short of telling us when certain subcircuits are engaged or how they give rise to particular behaviours.

Encouragingly, efforts have been made to bridge this gap, e.g., by way of either causally fitting parametric latent circuits to neural data (Langdon & Engel, 2025) or reverse-engineering canonical dynamical motifs in trained networks (Torre et al., 2025). However, these methods typically require assuming a circuit architecture a priori or focus on very specific dynamic regimes rather than providing general-purpose methods for broad interrogation of learnt circuits. While these approaches can be effective in confirmatory settings when a candidate circuit or motif is already defined and the goal is to test a specific mechanistic hypothesis, they are less well suited to exploratory circuit discovery where one must first identify the relevant subcircuits and generate hypotheses from scratch.

On the other hand, a growing body of work in "mechanistic interpretability" (Saphra & Wiegreffe, 2024; Rai et al., 2024) has developed tools to address this challenge in large language models (LLMs) (Sharkey et al., 2025; Anthropic, 2025) through the lens of causal abstraction (Geiger et al., 2020; 2021). Leveraging techniques such as ablations and activation patching to perturb internal activations and measure the causal effect on model output to score their importance (Vig et al., 2020; Geiger et al., 2021; Meng et al., 2022; Heimersheim & Nanda, 2024), these interventions can be applied at multiple granularities within a deep neural network, from individual neurons to higher-level computational subgraphs. Such techniques have enabled researchers to localize where specific computations occur, reverse-engineer multi-component circuits responsible for complex behaviors (Wang et al., 2022), and develop principled strategies for identifying task-critical components (Conmy et al., 2023; Zhang & Nanda, 2023; Li & Janson, 2024), hence establishing causal interventions as a methodological backbone for linking network components to computational function.

Yet, these methods have been developed almost exclusively for feedforward architectures where information flows through distinct, sequentially ordered layers; Adapting them to RNNs, where the effects of shared parameters compound across timesteps[2] through the evolving hidden state, requires tools that can resolve circuit dynamics across the temporal unfolding of a computation.

[2]Indeed, the compounding here is not just weight reuse across parallel positions of a fixed-depth graph but the recurrent composition of one nonlinear map through successive hidden activations, with no residual stream over which influences decompose linearly, requiring the ability to find circuits that are state-dependent rather than fixed.

1.2. Contributions

In this work, we present a pipeline for **circuit tracing in recurrent architectures** that allows us to **identify temporal, dynamically coordinated, neuron-indexed subcircuits**, thus providing **RNN interpretability in the form of explicit structure-function mappings**. Specifically, our contributions include:

- A principled approach for time-resolved neural circuit discovery in RNNs that shows *which* neurons matter, *when* they matter, and *how* they interact. We achieve this through a two-stage process combining i) targeted, windowed causal interventions on the neurons' hidden states to identify task-critical neurons at each timestep, with ii) Jacobian-based linearization of the recurrence in the network to expose the dynamic effective connectivity amongst them.
- Validation on synthetic tasks with known ground truth, where we demonstrate that our methods recover circuits in ring attractor networks and Hidden Markov Model (HMM) inference tasks with high precision, with our causally-selected neuronal subsets matching full-network decoder performance and maintaining robustness under noise.
- Recovering distinct, dynamically reconfiguring subcircuits for memory maintenance, state switching, and context-gated control that show division of excitatory/inhibitory (E/I) labour in RNNs respecting Dale's law, when trained to perform a context-dependent flip-flop task.
- Findings with RNNs trained on the Allen Institute's visual behaviour dataset (Garrett et al., 2023; Allen Institute for Brain Science, 2021) that suggest the intrinsic timing of VIP interneurons drives prediction error formation, bridging the timing-code and predictive-processing interpretations of VIP function.

2. Mechanistic Interpretability for Recurrent Computation

Consider the standard Elman RNN (Elman, 1990) update where the current hidden state h_t is determined by the hidden state at the previous time step h_{t-1} and the current input x_t, acting through the recurrent weights W_{hh} and input projection matrices W_{hi}, with b_{hh}, b_{hi} as their corresponding biases, and a non-linearity ϕ as follows:

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

Given that the same recurrent parameters are applied repeatedly along an evolving hidden-state trajectory, computations implemented by an RNN cannot in general be localized to

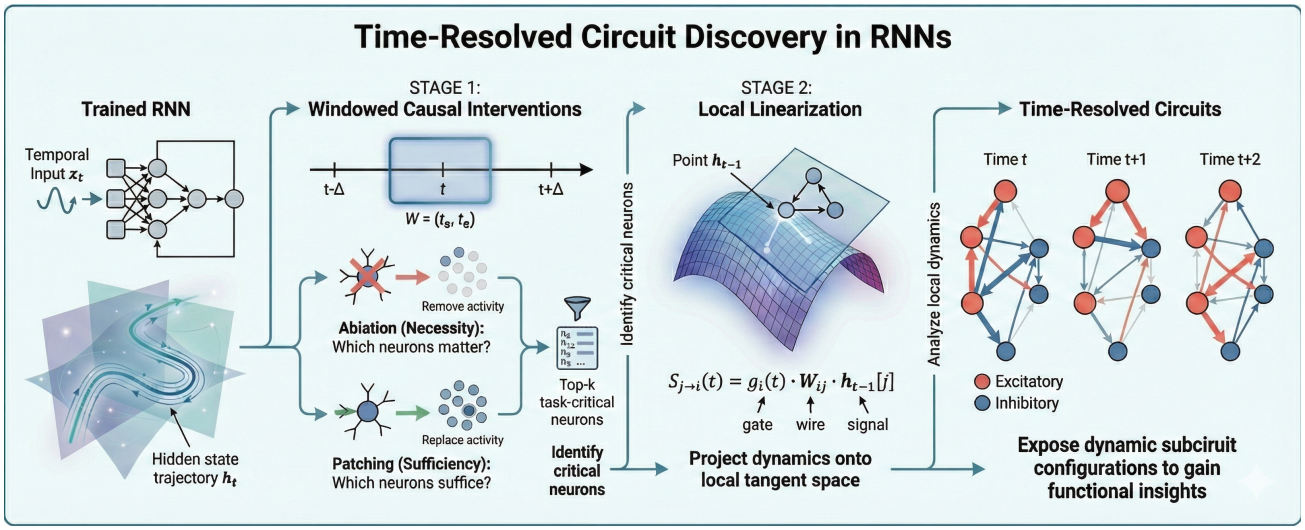

Figure 1. **Time-resolved neuronal circuit discovery pipeline for RNNs.** Given a trained RNN, our method proceeds in two stages. Windowed causal interventions (Stage 1) identifies task-critical neurons by intervening on hidden state activations within temporal windows $W = (t_s, t_e)$. Ablation tests necessity while activation patching tests sufficiency. This yields a ranked set of causally-important neurons for each computational epoch. Local linearization (Stage 2) characterizes how all neurons interact by computing the instantaneous effective connectivity $S_{j\to i}(t)$, which decomposes influence into postsynaptic gain (gate $g_i(t)$), synaptic weight (wire W_{ij}), and presynaptic activity (signal $h_{t-1}[j]$). This projects the recurrent dynamics onto the local tangent space at each timestep. Combining the two stages, the time-resolved circuit diagrams reveal how subcircuit configurations dynamically reconfigure across a computation.

fixed subgraphs of the recurrent weight matrix; its influence is state-dependent and may be amplified, attenuated, or overwritten across the rollout. This makes circuit discovery susceptible to spatiotemporal conflation, where activity at one timestep, correlation across a trajectory, or downstream consequences of a computation may be mistaken for components that causally support a behaviour of interest. Furthermore, since the recurrent dynamics are nonlinear, the operative interactions also cannot be read directly from the static weight matrix, but must instead be recovered through a local approximation of the dynamics. Finally, circuit traces across time require components with stable and biologically meaningful identities, so that computational roles can be ascribed to elements of the model in a way that preserves the structure afforded by the neuronal state space, rather than to abstractions induced by transformations of the hidden activity.

We therefore codify a set of assumptions underlying our framework, which make precise what counts as an operative circuit, how its components are implicated, the basis in which they are represented, and how their effective interactions are recovered.

(A1) **Circuits are dynamic, state-dependent configurations:** Even with fixed synaptic weights, the pattern of effective influence in RNNs varies with the current network state. Accordingly, the operative circuit at time t is not the weight matrix itself, but instead the *state-conditioned* map of those connections which are currently functionally active, and how strongly they transmit signals at that moment.

(A2) **Computational roles of circuit elements are causally identifiable via targeted interventions:** Which neurons contribute to what behaviour-relevant computations and when, can be identified by intervening on their activations and measuring the resulting change in the model output. Windowed interventions localize these roles in time, resolving transient contributions that would otherwise be conflated or lost using coarse or static perturbations.

(A3) **Neurons themselves provide a privileged basis for temporal circuits:** Unlike feedforward networks where a neuron might be permanently assigned to detecting a single feature, recurrent neurons participate dynamically in computation: the same neuron can belong to one circuit at timestep t and switch to another at timestep $t + 1$, or be quiescent during periods when its function is not required. Circuit membership is therefore not static: a neuron's computational role is determined by when it is recruited and how it interacts with other neurons in the network at each timestep, making the neuronal basis privileged and non-superposed[3] for understanding computation in RNNs.

[3]Recent work also provides independent support for the privileged neuronal basis assumption in LLMs (Arora et al., 2026), demonstrating that MLP neurons in transformers are in fact as sparse a feature basis as dictionaries learnt with sparse autoencoders. They attribute this to the element-wise nonlinearity, which

(A4) **Local linearization suffices for circuit characterization:** The effective circuit at each timestep is well-characterized by the local linear approximation of the recurrent dynamics. The hidden-state Jacobian, weighted by current pre-synaptic activity, captures which connections are actively transmitting influence, sufficient to reveal circuit structure without requiring higher-order interaction terms.

It therefore follows from our postulates above that: *If biological circuits implement canonical computations, and if RNNs trained on appropriate objectives converge to similar solutions, then the circuits we identify should reflect genuine computational primitives rather than idiosyncratic network solutions or artifacts.*

In the following subsections, we describe the two steps of our pipeline that help us turn trained RNNs into time-resolved circuit diagrams (Fig. 1). First, windowed causal interventions allow us to identify *which* neurons are task-critical during specific temporal windows. Next, time-resolved linearization of the recurrent dynamics via the Jacobian reveals *how* these neurons interact at each timestep.

2.1. Time-Windowed Causal Interventions

To identify neurons causally responsible for task performance within specific time windows, we perform targeted interventions on hidden unit activity and measure changes in behavioral output.

Given a temporal window $w = [t_s, t_e)$ aligned to a computational epoch of interest, we intervene on a neuron's hidden state trace within w and then re-run the forward pass from t_s onward, allowing downstream effects to propagate naturally through the recurrence.

The effect of each intervention is quantified via the per-timestep decision margin[4] where for output logits $z_t \in \mathbb{R}^C$ and correct label $y_t^\star$, the margin is

$$m_t(z_t, y_t^\star) = z_t[y_t^\star] - \max_{c \neq y_t^\star} z_t[c].$$

We subsequently compute causal importance scores using either or both of two complementary interventions as appropriate, one of which tests necessity and the other sufficiency.

Soft ablation (necessity). To test whether a set of neurons I is necessary for task performance in the window w, we attenuate their hidden states by a factor $\alpha \in [0, 1)$ within the window and re-roll the dynamics forward:

$$h_{t,i}^{\text{abl}} = \begin{cases} \alpha\, h_{t,i}, & i \in I,\ t \in w, \\ h_{t,i}, & \text{otherwise.} \end{cases}$$

The ablation score is the average margin drop:

$$\Delta_{\text{abl}}(I) = \frac{1}{|w|} \sum_{t \in w} \big(m_t - \tilde{m}_t\big),$$

where $\tilde{m}_t$ is the margin after ablation. A positive Δ_{abl} indicates that the ablated neurons are necessary for maintaining task performance within w.

Trace patching (sufficiency). To test whether a single neuron $i^\star$ is sufficient to drive a particular computation, we copy its hidden state trace from a *source* trial (e.g., one containing a state switch) into a *target* trial (e.g., a hold trial) within w, and re-roll the target dynamics:

$$h_{t,i^\star}^{\text{patch}} \leftarrow h_{t,i^\star}^{\text{src}}, \quad t \in w.$$

The patching score is:

$$\Delta_{\text{patch}}(i^\star) = \frac{1}{|w|} \sum_{t \in w} \big(\tilde{m}_t - m_t\big).$$

A positive Δ_{patch} indicates that transplanting this neuron's activity is sufficient to push the target trial's output toward the source trial's behavior, where we expect the margin after patching, $\tilde{m}_t$ to increase if the neuron carries the relevant computation.

From scores to subcircuits. For each computational epoch defined by a window w, we select the top-k neurons by causal importance score to define a functionally specialized subcircuit. Crucially, this set can (and typically does) change across windows: a neuron that is critical during one epoch may become quiescent or even counterproductive during a subsequent one, and vice versa.

The window size itself controls the temporal resolution of the analysis. At one extreme, setting w equal to the full rollout recovers a single, time-averaged importance ranking akin to conventional ablation studies, though notably still distinct from static weight magnitude comparisons, as the ablation score remains activity-weighted and captures indirect effects propagated through the recurrence. At the other extreme, single-timestep windows ($|w| = 1$) resolve the finest-grained snapshot of which neurons are operative at each moment.

Intermediate window sizes offer a practical middle ground, aggregating importance over a computationally coherent

breaks rotational symmetry and makes individual neurons structurally privileged. The same argument could apply to RNNs, with the activation function ϕ acting as the pointwise nonlinearity enabling a privileged coordinate system in neuronal space.

[4]For RNNs not trained on classification tasks, where output logits do not directly correspond to class labels, the margin can instead be defined with respect to per-timestep linear readouts trained post hoc on the hidden state population (e.g., logistic regression classifiers fit to discriminate task-relevant conditions at each timestep, with $m_t = w_t^\top h_t + b_t$ oriented so that a positive margin indicates a correct classification).

epoch while remaining sensitive to the transient recruitment and release of individual neurons. This choice also governs computational cost: each intervention requires a forward re-roll from t_s to the end of the sequence, so the total number of re-rolls scales as $\mathcal{O}(N \cdot T/|w|)$ where N is the number of neurons evaluated and T is the sequence length, making finer windows linearly more expensive.

2.2. Recurrence Linearization

While causal interventions identify *which* neurons are important, time-resolved linearization reveals *how* they interact. We derive a decomposition of the instantaneous effective connectivity in the RNN that factorizes each neuron-to-neuron influence into three interpretable components.

Recall the Elman RNN update from Eq. 1, and let $a_t = W_{hi}x_t + b_{hi} + W_{hh}h_{t-1} + b_{hh}$ denote the pre-activation. Since ϕ is applied elementwise, the hidden-to-hidden Jacobian decomposes as

$$J_t^{(h)} = \frac{\partial h_t}{\partial h_{t-1}} = D_{\phi,t}\, W_{hh}$$

where $D_{\phi,t} = \mathrm{diag}(\phi'(a_t))$ is the diagonal matrix of pointwise derivatives evaluated at the current pre-activation. We refer to $D_{\phi,t}$ as the *postsynaptic gain matrix*, since its i-th diagonal entry $g_i(t) = \phi'(a_t[i])$ determines how sensitive neuron i is to incoming signals at time t.

We score the *instantaneous recurrent influence* from presynaptic neuron j to postsynaptic neuron i at time t by weighting the (i, j)-th entry of the Jacobian by the presynaptic activity:

$$S_{j\to i}(t) = J_t^{(h)}[i,j] \cdot h_{t-1}[j] = \underbrace{g_i(t)}_{\text{gate}} \cdot \underbrace{W_{hh}[i,j]}_{\text{wire}} \cdot \underbrace{h_{t-1}[j]}_{\text{signal}}$$

Interpreting the influence score. The gate–wire–signal factorization has a natural interpretation: an edge carries influence only if the postsynaptic neuron is sensitive (i.e., gate is open), the connection exists (i.e., the wire is present), and the presynaptic neuron is active (i.e., a signal is sent).

Geometrically, the Jacobian $J_t^{(h)}$ projects the recurrent dynamics onto the local tangent space at each timestep; multiplication with the presynaptic state then selects which of those linearized pathways actually transmit signal at that moment, yielding a first-order estimate of how much each presynaptic neuron individually contributes to each postsynaptic neuron's state update.

Alternatively, one could also interpret the influence score as the gradient $\times$ activation *attribution* of the one-step recurrence: by multiplying the Jacobian entry $\frac{\partial h_t[i]}{\partial h_{t-1}[j]}$ with the presynaptic activity $h_{t-1}[j]$, we attribute the postsynaptic update $h_t[i]$ to its presynaptic inputs. In this sense, the linearization is similar to first-order, gradient-based methods like attribution patching (Nanda, 2023; Syed et al., 2024).

Moreover, integrated gradients (IG) as defined by Sundararajan et al. (2017) taken with a zero baseline, returns the path-averaged form of $S_{j\to i}(t)$ which reports the instantaneous value at the current state. The two are exactly equal in the case of the recurrence being locally affine along the path (e.g., if ϕ is affine over the segment traversed, as for a ReLU unit that does not cross its threshold along the path), and $S_{j\to i}(t)$ is the leading term of IG otherwise. Furthermore, our choice of the single-point evaluation is deliberate: Given that $S_{j\to i}(t)$ and IG differ only in the gate, i.e., the instantaneous gain $g_i(t)$ for $S_{j\to i}(t)$ vs the gain averaged over the interpolated states for IG, our score preserves the transient, state-conditioned gating that distinguishes one operative circuit from the next, whereas IG's averaging obscures the same. We refer the reader to Appendix A for the derivation of the exact relationship between IG and our score.

Lastly, when ϕ is monotonically non-decreasing (as is the case for standard choices such as $\tanh$, ReLU, and their compositions) we have $g_i(t) \geq 0$ for all i and t. The sign of $S_{j\to i}(t)$ is therefore determined entirely by the wire and signal terms. In networks obeying Dale's law, where $\mathrm{sign}(W_{hh}[i,j])$ is fixed by cell type, this makes the sign of each edge directly interpretable as excitatory or inhibitory.

3. Toy Models of RNN Circuitry

Having established the methodological framework in Sec. 2, we evaluate our pipeline[5] in two settings of increasing complexity: synthetic benchmarks with known ground truth (Sec. 3.1), followed by an RNN respecting Dale's law that is trained on a context-dependent flip-flop task, that lets us study how excitatory and inhibitory neurons coordinate across distinct computational demands (Sec. 3.2).

3.1. Circuit Recovery on Synthetic Benchmarks

Ring attractor network. We construct a 24-neuron network (19 excitatory, 5 inhibitory) implementing a ring attractor that encodes four spatial positions through localized excitatory clusters (e.g., neurons 12–15 encode position 3). At test time, we generate three trajectory types that probe circuit discovery under qualitatively different dynamics: i) *static bumps* that maintain fixed positions, ii) continuously drifting *traveling bumps*, and iii) *jumping bumps* with discrete transitions between non-adjacent positions on the ring.

Windowed causal interventions (top-k=5 neurons) recover

[5] All computational experiments throughout the paper were carried out on a Dell Pro Max 14 Premium with an Intel Core Ultra 7 processor, 32 GB RAM, and an Intel Arc Pro 140T 16 GB GPU.

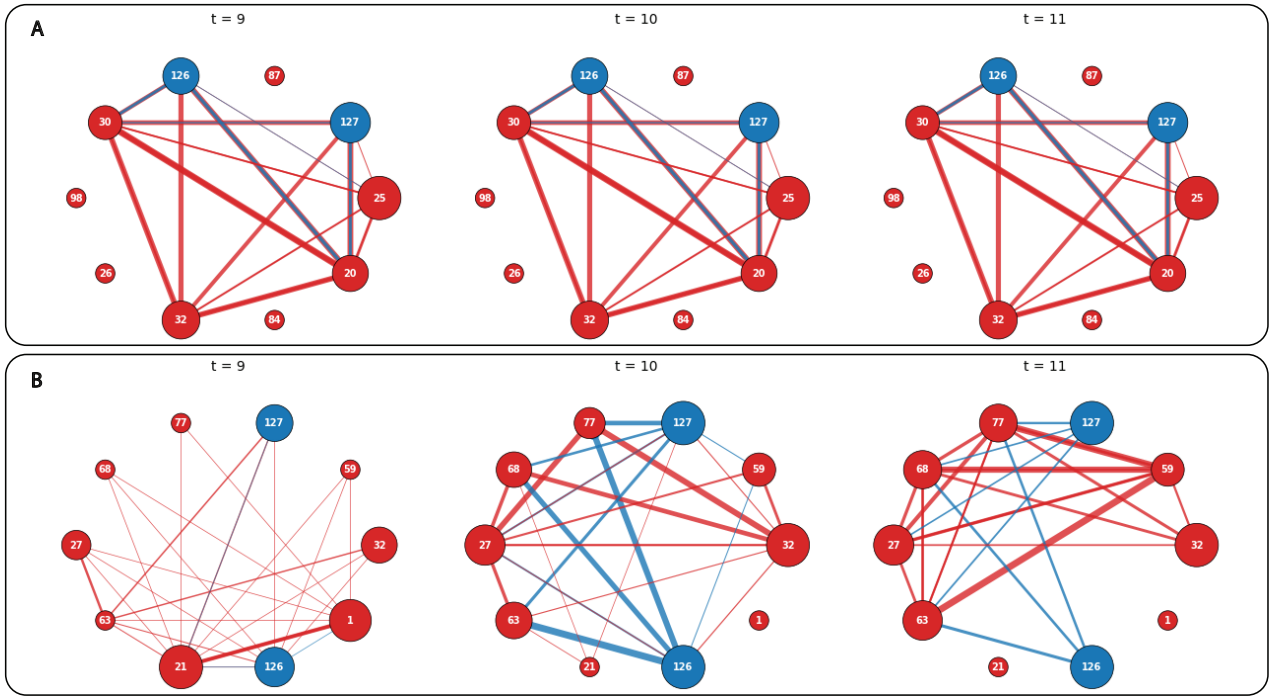

Figure 2. **Time-resolved circuits for the context-dependent flip-flop task.** (A) Memory and (B) cue-switching circuits at timesteps t=9, 10, 11 around the switching event at t=10. Only the top-10 neurons (by windowed causal score) are shown. Node color: excitatory (red), inhibitory (blue). Node size: proportional to normalized hidden activity at time t. Edge color: sign of effective influence. Edge width: magnitude of instantaneous influence $|S_{j\to i}(t)|$. Top row: The memory circuit remains stable across all three timesteps. Bottom row: The cue-switching circuit transiently recruits inhibitory gating at t=10, i.e., at the switching event.

ground truth position-encoding clusters with high precision, achieving mean F1 of 91% across static positions, and 92% on jumping bumps. Performance on traveling bumps is lower (68% F1), reflecting the inherent ambiguity of the ground truth "important set" when the bump is in continuous motion. Averaging across all trajectory types and positions yields a mean F1 of 83.67% (Appendix B, Top row, Table 1). Detailed per-position F1 breakdowns are provided in the Appendix (Table 2).

We also find that the time-resolved circuit diagrams produced by our pipeline clearly reveal the dynamic reorganization during transitions; neurons encoding the initial position dominate causal importance before the jump, while those encoding the target position become critical after (Appendix C). We correctly identify causally relevant neurons at each timestep via windowed interventions and expose how their effective interactions reconfigure through the linearized influence scores, recovering stable recurrent loops for static bumps (Fig. 4), transient inhibitory gating during discrete jumps (Fig. 5), and progressive handoff of excitatory drive in traveling bumps (Fig. 6).

HMM inference task. We train a 128-neuron RNN to predict the next observation in sequences generated from a 3-state hidden Markov model with diagonal-dominant transition dynamics and state-specific emission patterns. The network implicitly learns to infer hidden states from observation history; ground truth state beliefs are computed via Baum-Welch forward inference (Rabiner, 2002).

Applying windowed ablation during sequences that remain in a single state for extended periods, we identify state-specific maintenance circuits, 5 neurons per state (15 total). These discovered circuits achieve functional sufficiency, matching full-network classification accuracy (100%) with an 8.5× parameter reduction (Appendix B, Bottom row, Table 1). Per-state binary classification confirms that each 5-neuron circuit specifically encodes its target state, significantly outperforming random neuron selection (Appendix D, Fig. 8). Discovered neurons also exhibit sharp temporal specificity, with high causal importance precisely during their state's maintenance window and near-zero importance elsewhere (Appendix D, Fig. 7). To test whether these circuits capture robust computational structure rather than mere correlations, we evaluated performance under additive Gaussian noise ($\sigma \in [0, 0.5]$) added to hidden states during decoding. Across all noise levels, discovered circuits maintained average accuracies of 77.7%, outperforming correlation-based Baum-Welch neuron selection (68.3%; Appendix D, Table 3).

Together, these results validate our pipeline's capabilities for precise discovery of causal subcircuits that are functionally sufficient and robust under noise.

3.2. E/I Motifs in Context-Dependent Flip-Flop RNNs

Having validated our methods, we apply them to an RNN trained to perform a contextually-gated flip-flop task with E/I constraints, requiring our pipeline to be able to disentangle three computational demands commonly co-occurring in cortical settings: memory maintenance, state switching, and context-gated control and therefore shed light on how excitatory and inhibitory neurons are differentially recruited across these demands.

Task. We study a context-dependent flip-flop task (Sussillo & Barak, 2013) in which the network must maintain a latent state $y_t \in \{-1, 0, 1\}$ and update it only when a relevant input pulse arrives. At each timestep t, the input x_t contains two pulse channels (A, B) and a context bit $c_t \in \{\mathrm{A}, \mathrm{B}\}$ indicating which channel is currently relevant; distractor pulses may occur on the irrelevant channel. An in-context "set" pulse switches the state to $+1$, an in-context "reset" pulse switches it to -1, and in the absence of a relevant pulse the state persists. Context can change mid-sequence, after which relevance swaps across channels and the output should immediately reflect the correct state under the new context. The general task structure is illustrated in Appendix E (Fig. 9a).

Model. We use a single-layer, Dale-constrained Elman RNN with a $\tanh$-then-rectify hidden nonlinearity, trained via cross-entropy with Dale's backpropagation (Balwani et al., 2025b) which enforces Dale's law by projecting W_{hh} onto the appropriate sign-constrained orthant after each gradient update. Hidden units are partitioned into excitatory and inhibitory populations in an 8:2 ratio, with 128 hidden neurons in total. Full architectural and training details are provided in Appendix F.

Diagnostic stimuli. To isolate the computational demands of memory, cue-switching, and state-switching and identify their associated subcircuits, we utilize diagnostic inputs from three controlled stimulus families designed to engage a single computational demand in isolation:

- **Memory probe:** Initialize the state with a single in-context pulse, then present a long delay with either no pulses or irrelevant distractors only (Appendix E, Fig. 9b).
- **State-switching probe:** In-context, deliver a pulse that flips the state (set→reset or reset→set) with no context change and no distractors (Appendix E, Fig. 9c).
- **Cue-switching probe:** Establish a state, then flip the context channel while injecting distractors on the now-irrelevant channel and occasional relevant pulses on the newly relevant channel (Appendix E, Fig. 9d).

Results. The *memory circuit* (Fig. 2A) comprises a stable excitatory cluster with persistent node sizes and recurrent edges from t=9 to t=11, while inhibitory nodes show sparse connectivity consistent with modulatory roles. In contrast, the *cue-switching circuit* (Fig. 2B) exhibits a striking temporal dissociation: the excitatory cluster remains stable, but inhibitory nodes transiently expand at t=10 with thickened edges to excitatory targets, then subside by t=11. This reveals coordinated E/I specialization – stable excitatory loops sustain memory while transient inhibitory gating enables context control without disrupting storage. *State-switching* (Appendix G) displays similar but weaker, delayed inhibitory dynamics, reflecting the need to overcome stable excitatory maintenance to drive the state flip. In toto, the takeaway remains: recurrent excitation dominates memory, while switching recruits inhibitory neurons at transition points, and context switches are mediated by a gating circuit whose connectivity reconfigures rapidly while the memory circuit remains stable. This temporal dissociation – stable maintenance coexisting with transient gating – would be invisible to static summaries of weights or activity, and illustrates the value of time-resolved circuit analysis.

4. Bridging Notions of Timing and Predictive Processing in Mouse V1

We now apply our pipeline to an anatomically-constrained RNN trained on neural data from mouse primary visual cortex, where the ground truth circuit is unknown and the goal is to generate mechanistic hypotheses about cell-type-specific computation. All results are averaged across 5 runs.

Dataset and model. We analyze a Dale's law-constrained RNN trained on the Allen Institute Visual Behavior Two-Photon dataset (Garrett et al., 2023; Allen Institute for Brain Science, 2021), containing calcium imaging from Layers 2/3 (L2/3) of mouse primary visual cortex (V1) where the cortical circuit comprises three genetically identified cell classes: excitatory pyramidal (Pyr) neurons and two inhibitory interneuron types, somatostatin-expressing (SST) and vasoactive-intestinal-peptide-expressing (VIP). Cosequently, Pyr neurons have non-negative outgoing weights, while SST and VIP neurons have non-positive outgoing weights.

The model follows a restricted V1 L2/3 version of the Cell-typeRNN from Balwani et al. (2025b) and contains 240 neurons (192 Pyr, 32 SST, 16 VIP) with sparse inter-population connectivity constrained by the synaptic wiring patterns

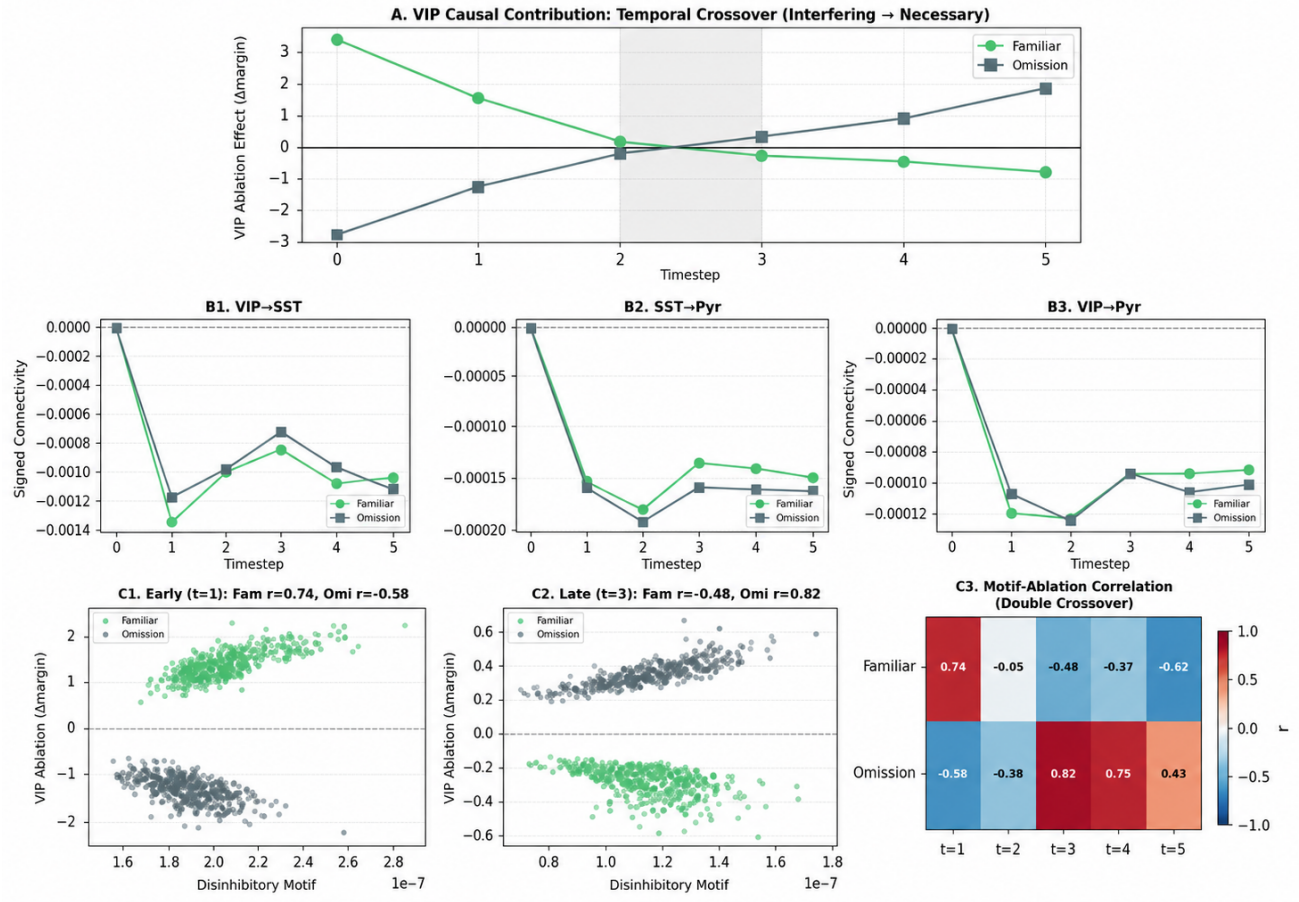

Figure 3. **Time-resolved VIP–SST–Pyr circuit reconfiguration during stimulus omission. A.** VIP causal contribution (Δmargin upon ablation) for Familiar (green) and Omission (grey) trials across 6 timesteps. A temporal crossover occurs at $t\approx 2$–3: VIP ablation helps omission classification early but hurts it late. Grey shading marks the crossover window. **B1–B3.** Signed effective connectivity for the three key inhibitory pathways (VIP→SST, SST→Pyr, VIP→Pyr) computed via recurrence linearization, showing condition-dependent divergence aligned with the causal crossover. **C1–C2.** Trial-by-trial scatter plots of disinhibitory motif strength (VIP→SST×SST→Pyr) vs. VIP ablation importance at an early (t=1) and late (t=3) timestep, demonstrating that TRL motif strength predicts causal importance with sign-flipping correlations across conditions and time. **C3.** Motif–ablation Pearson r for both conditions across all timesteps, revealing a double crossover: the same circuit motif transitions from predicting interference to predicting necessity.

measured by Campagnola et al. (2022) (Table 4, Appendix I) where each cell type receives external input only from its corresponding channel. Every cell-type specific population's activity is projected to per-population outputs through a randomly initialized linear readout layer that is held fixed throughout training, so that condition-dependent differences in the inferred circuit reflect changes in the recurrent dynamics rather than a re-tuned output mapping.

The recordings themselves come from mice performing a go/no-go visual change-detection task, viewing a stream of briefly flashed natural images and reporting changes in image identity; on a small fraction of trials the expected image is omitted and replaced by a blank screen, a controlled violation of the animal's learned expectation that makes the paradigm a natural testbed for predictive processing. The task contrasts *Familiar* trials (expected stimulus appears) with *Omission* trials (expected stimulus withheld) across 6 timesteps ($\times$125ms = 750ms post-stimulus window). Additional details about the behavioural task and training data are provided in Appendix H.

The model is trained to perform one-step-ahead prediction of neural responses, minimizing mean squared reconstruction error[6] computed independently within each cell-type population and averaged across the three. Full model architectural and training details appear in Appendix I.

A temporal crossover in VIP causal importance. Applying windowed ablation independently per cell type (Ap-

[6]Since the network is not trained to classify conditions directly, the decision margin for our causal interventions is defined via per-timestep logistic regression classifiers fit post hoc on the hidden states to discriminate Familiar from Omission trials (see Sec. 2.1).

pendix J, Fig. 11), we find that Pyr and SST populations show relatively uniform importance profiles across time, with no striking condition×time interaction. VIP neurons, by contrast, exhibit a pronounced temporal crossover (Fig. 3A): at early timesteps (t=0–2), VIP ablation *improves* omission classification, i.e., VIP activity actively interferes with correct discrimination, while at late timesteps (t=3–5), VIP ablation *impairs* it, revealing that VIP has become necessary for detecting the unexpected stimulus absence. For familiar trials the pattern reverses: VIP is causally important early but dispensable late. This crossover, invisible to any time-averaged importance measure (Appendix L shows that population-level correlational measures such as activity difference and classifier weights miss this crossover entirely), occurs around t=2–3 (∼300–400ms post-stimulus). Recurrence linearization of the three key inhibitory pathways (Fig. 3B1–B3) corroborates this picture: the VIP→SST→Pyr disinhibitory pathway is weaker for omission trials early but stronger late, with condition-dependent divergence emerging precisely at the crossover. Appendix M shows the complete set of signed effective connectivities across all permutations of the {Pyr, SST, VIP} populations.

Unifying perspectives on the role of VIP neurons for predictive processing. These results bridge two influential accounts of VIP function that make distinct predictions in this setting. The predictive-processing account (Najafi et al., 2025) posits that VIP neurons signal prediction errors, predicting VIP necessity for omission detection throughout the trial. On the other hand, the interval-timing account (Huang et al., 2025) proposes that VIP ramps reflect elapsed time since the last stimulus, predicting a crossover in causal importance as the timing signal is reinterpreted by downstream circuits (see Appendix K for the corresponding hidden-state ramping dynamics). VIP ablation impairs omission classification at late timesteps, consistent with (Najafi et al., 2025), but the early-phase inversion, where VIP ablation *helps*, aligns instead with (Huang et al., 2025), suggesting that early VIP activity reflects a timing signal not yet reinterpreted as surprise. Trial-by-trial motif–ablation correlations (Fig. 3C) reinforce this interpretation with a double crossover: the VIP→SST→Pyr motif positively predicts VIP importance on familiar trials early (r=0.74) but negatively on omission (r=−0.58), with both relationships inverting by late timesteps – indicating that VIP's functional *mode*, not merely its activity, transitions over time (full trial-by-trial pathway analysis provided in Appendix N). This suggests that VIP's intrinsic timing dynamics provide the substrate upon which prediction error signals are constructed, with the ∼300–400ms reconfiguration timescale constituting a testable prediction for temporally-targeted VIP silencing experiments (Appendix O).

5. Discussion

By combining windowed causal interventions with recurrence linearization, our pipeline turns trained RNNs into time-resolved circuit diagrams that expose which neurons matter, when they matter, and how they interact, producing the explicit structure–function maps needed to generate testable hypotheses about neural computation as evidenced by our recovery of ground truth circuits in synthetic benchmarks, discovery of coordinated E/I specialization in a flip-flop task, and identification of a temporally-targeted VIP silencing prediction in mouse V1, illustrating how mechanistic interpretability of RNNs might hold much promise for closing the loop between computational modelling, hypothesis generation, and biological experiments.

Yet, several limitations point to future work. The recurrence linearization is first-order and may miss cooperative nonlinear interactions; extending to second-order terms could capture effects the current factorization treats independently. Window size governs temporal resolution and must be chosen manually; adaptive strategies that detect transition points would reduce this degree of freedom. Our V1 analysis uses a single architecture restricted to L2/3; multi-region, multi-layer RNNs could test whether the discovered motifs generalize across cortical hierarchies and also interrogate questions of greater complexity than our initial attempt.

Impact Statement

This work aims to advance both ML interpretability and computational neuroscience by demonstrating that causal circuit discovery methods can be fruitfully applied to RNNs trained to replicate the dynamics of neural data. By finding task and behaviour-relevant circuits in RNNs, we believe our approach could in the future contribute to understanding dysfunction in neurological and psychiatric conditions. More broadly, we also anticipate that insights gained from identifying and interpreting circuits in biologically-constrained models could inform the design of more efficient or robust architectures in conventional ML. At present, we do not foresee specific negative societal consequences arising from this work beyond those generally associated with bad actors misusing the knowledge gained through advances in mechanistic interpretability.

References

Allen Institute for Brain Science. Allen Brain Observatory: Visual Behavior 2-Photon Calcium Imaging Dataset. Allen Brain Map / Brain Knowledge Platform, 2021. URL https://portal.brain-map.org/explore/circuits/visual-behavior-2p. Accessed May 5, 2026.

Anthropic. Transformer circuits thread, 2025. URL https://transformer-circuits.pub/. [Online; accessed 2026-01-27].

Arora, A., Wu, Z., Steinhardt, J., and Schwettmann, S. Language model circuits are sparse in the neuron basis. *arXiv preprint arXiv:2601.22594*, 2026.

Balwani, A., Cho, S., and Choi, H. Exploring the architectural biases of the cortical microcircuit. *Neural Computation*, 37(9):1551–1599, 2025a.

Balwani, A., Wang, A. Q., Najafi, F., and Choi, H. Constructing biologically constrained rnns via dale's backpropagation and topologically informed pruning. *Science Advances*, 11(50):eadw4970, 2025b.

Barak, O. Recurrent neural networks as versatile tools of neuroscience research. *Current opinion in neurobiology*, 46:1–6, 2017.

Campagnola, L., Seeman, S. C., Chartrand, T., Kim, L., Hoggarth, A., Gamlin, C., Ito, S., Trinh, J., Davoudian, P., Radaelli, C., et al. Local connectivity and synaptic dynamics in mouse and human neocortex. *Science*, 375 (6585):eabj5861, 2022.

Cao, R. and Yamins, D. Explanatory models in neuroscience: Part 2–constraint-based intelligibility. *arXiv preprint arXiv:2104.01489*, 2021.

Cao, R. and Yamins, D. Explanatory models in neuroscience, part 1: Taking mechanistic abstraction seriously. *Cognitive Systems Research*, 87:101244, 2024.

Conmy, A., Mavor-Parker, A., Lynch, A., Heimersheim, S., and Garriga-Alonso, A. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023.

DePasquale, B., Cueva, C. J., Rajan, K., Escola, G. S., and Abbott, L. full-force: A target-based method for training recurrent networks. *PloS one*, 13(2):e0191527, 2018.

Douglas, R. J. and Martin, K. A. Neuronal circuits of the neocortex. *Annu. Rev. Neurosci.*, 27(1):419–451, 2004.

Douglas, R. J. and Martin, K. A. Recurrent neuronal circuits in the neocortex. *Current biology*, 17(13):R496–R500, 2007.

Douglas, R. J., Koch, C., Mahowald, M., Martin, K. A., and Suarez, H. H. Recurrent excitation in neocortical circuits. *Science*, 269(5226):981–985, 1995.

Driscoll, L. N., Shenoy, K., and Sussillo, D. Flexible multitask computation in recurrent networks utilizes shared dynamical motifs. *Nature Neuroscience*, 27(7):1349–1363, 2024.

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

Garrett, M., Groblewski, P., Piet, A., Ollerenshaw, D., Najafi, F., Yavorska, I., Amster, A., Bennett, C., Buice, M., Caldejon, S., et al. Stimulus novelty uncovers coding diversity in visual cortical circuits. *BioRxiv*, pp. 2023–02, 2023.

Geiger, A., Richardson, K., and Potts, C. Neural natural language inference models partially embed theories of lexical entailment and negation. *arXiv preprint arXiv:2004.14623*, 2020.

Geiger, A., Lu, H., Icard, T., and Potts, C. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34:9574–9586, 2021.

Heimersheim, S. and Nanda, N. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024.

Huang, Y., Shamsnia, A., Chen, M., Wu, S., Stamm, T., Medico, S., and Najafi, F. Intrinsic timing, not temporal prediction, underlies ramping dynamics in visual and parietal cortex during passive behavior. *bioRxiv*, pp. 2025–09, 2025.

Langdon, C. and Engel, T. A. Latent circuit inference from heterogeneous neural responses during cognitive tasks. *Nature neuroscience*, 28(3):665–675, 2025.

Li, M. and Janson, L. Optimal ablation for interpretability. *Advances in Neural Information Processing Systems*, 37: 109233–109282, 2024.

Li, P., Cornford, J., Ghosh, A., and Richards, B. Learning better with dale's law: A spectral perspective. *Advances in Neural Information Processing Systems*, 36:944–956, 2023.

Maheswaranathan, N., Williams, A., Golub, M., Ganguli, S., and Sussillo, D. Reverse engineering recurrent networks for sentiment classification reveals line attractor dynamics. *Advances in neural information processing systems*, 32, 2019.

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

Mastrogiuseppe, F. and Ostojic, S. Linking connectivity, dynamics, and computations in low-rank recurrent neural networks. *Neuron*, 99(3):609–623, 2018.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Najafi, F., Russo, S., and Lecoq, J. Unexpected events trigger task-independent signaling in vip and excitatory neurons of mouse visual cortex. *Iscience*, 28(2), 2025.

Nanda, N. Attribution patching: Activation patching at industrial scale. *URL: https://www. neelnanda. io/mechanistic-interpretability/attribution-patching*, 15: 19, 2023.

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

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

Rabiner, L. R. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 2002.

Rai, D., Zhou, Y., Feng, S., Saparov, A., and Yao, Z. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*, 2024.

Rajan, K., Harvey, C. D., and Tank, D. W. Recurrent network models of sequence generation and memory. *Neuron*, 90(1):128–142, 2016.

Recanatesi, S., Farrell, M., Lajoie, G., Deneve, S., Rigotti, M., and Shea-Brown, E. Predictive learning as a network mechanism for extracting low-dimensional latent space representations. *Nature communications*, 12(1):1417, 2021.

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., et al. A deep learning framework for neuroscience. *Nature neuroscience*, 22(11): 1761–1770, 2019.

Russo, A. A., Khajeh, R., Bittner, S. R., Perkins, S. M., Cunningham, J. P., Abbott, L. F., and Churchland, M. M. Neural trajectories in the supplementary motor area and motor cortex exhibit distinct geometries, compatible with different classes of computation. *Neuron*, 107(4):745–758, 2020.

Saphra, N. and Wiegreffe, S. Mechanistic? *arXiv preprint arXiv:2410.09087*, 2024.

Saxe, A., Nelli, S., and Summerfield, C. If deep learning is the answer, what is the question? *Nature Reviews Neuroscience*, 22(1):55–67, 2021.

Schneider, S., Laiz, R. G., Filippova, A., Frey, M., and Mathis, M. W. Time-series attribution maps with regularized contrastive learning. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025. URL https://openreview.net/forum?id=aGrCXoTB4P.

Sharkey, L., Chughtai, B., Batson, J., Lindsey, J., Wu, J., Bushnaq, L., Goldowsky-Dill, N., Heimersheim, S., Ortega, A., Bloom, J., et al. Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496*, 2025.

Song, H. F., Yang, G. R., and Wang, X.-J. Training excitatory-inhibitory recurrent neural networks for cognitive tasks: a simple and flexible framework. *PLoS computational biology*, 12(2):e1004792, 2016.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Sussillo, D. and Abbott, L. F. Generating coherent patterns of activity from chaotic neural networks. *Neuron*, 63(4): 544–557, 2009.

Sussillo, D. and Barak, O. Opening the black box: low-dimensional dynamics in high-dimensional recurrent neural networks. *Neural computation*, 25(3):626–649, 2013.

Sussillo, D., Churchland, M. M., Kaufman, M. T., and Shenoy, K. V. A neural network that finds a naturalistic solution for the production of muscle activity. *Nature neuroscience*, 18(7):1025–1033, 2015.

Syed, A., Rager, C., and Conmy, A. Attribution patching outperforms automated circuit discovery. In *The 7th BlackboxNLP Workshop*, 2024. URL https://openreview.net/forum?id=RysbaxAnc6.

Torre, E., Viscione, M., Pompe, L., Grewe, B. F., and Mante, V. Mechanistic interpretability of RNNs emulating hidden markov models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=kn4QjJFZaN.

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., and Shieber, S. Investigating gender bias in

language models using causal mediation analysis. *Advances in neural information processing systems*, 33: 12388–12401, 2020.

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

Yang, G. R. and Wang, X.-J. Artificial neural networks for neuroscientists: a primer. *Neuron*, 107(6):1048–1070, 2020.

Zhang, F. and Nanda, N. Towards best practices of activation patching in language models: Metrics and methods. *arXiv preprint arXiv:2309.16042*, 2023.

A. Integrated Gradients as the Path-Average of the Instantaneous Recurrence Influence Score

We now make precise the relationship between the instantaneous influence score $S_{j\to i}(t)$ of Sec. 2.2 and integrated gradients (IG) as proposed by Sundararajan et al. (2017) by showing that IG, taken with a zero baseline, averages the influence score over the hidden states interpolating from the baseline to the current state, while $S_{j\to i}(t)$ is its value at the endpoint of that path, i.e., at the current state h_{t-1} itself.

Holding the input x_t and biases fixed, we treat the one-step RNN update as a function of the previous hidden state,

$$F_i(h) = \phi\big(a(h)\big)_i, \qquad a(h) = W_{hh}\, h + \big(W_{hi}x_t + b_{hi} + b_{hh}\big),$$

so that $F_i(h_{t-1}) = h_t[i]$ and the object being attributed is the presynaptic activity h_{t-1}.

IG attributes F_i relative to a baseline h^0 via

$$\mathrm{IG}_{j\to i} = \big(h_{t-1}[j] - h^0[j]\big) \int_0^1 \frac{\partial F_i}{\partial h_j}\bigg|_{h=h^0+\alpha(h_{t-1}-h^0)} d\alpha.$$

If we take the zero baseline, i.e., set $h^0 = 0$, the path reduces to $h = \alpha\, h_{t-1}$. Since F_i depends on h only through the pre-activation $a(h)$, and the path $h = \alpha h_{t-1}$ scales the recurrent contribution $W_{hh}h$ to $\alpha W_{hh}h_{t-1}$ while leaving the input-and-bias term $(W_{hi}x_t + b_{hi} + b_{hh})$ unchanged, the pre-activation along the path is

$$a^{(\alpha)} = \alpha\, W_{hh}h_{t-1} + \big(W_{hi}x_t + b_{hi} + b_{hh}\big), \qquad a^{(1)} = a_t,$$

giving $\partial F_i/\partial h_j\big|_{\alpha h_{t-1}} = \phi'\big(a^{(\alpha)}[i]\big)\, W_{hh}[i,j]$.

Substituting this derivative into the IG integral yields

$$\mathrm{IG}_{j\to i} = h_{t-1}[j] \int_0^1 \phi'\big(a^{(\alpha)}[i]\big)\, W_{hh}[i,j]\, d\alpha = W_{hh}[i,j]\, h_{t-1}[j] \int_0^1 \phi'\big(a^{(\alpha)}[i]\big)\, d\alpha$$

where we note that the integrand is the influence score with its gate evaluated at the *interpolated* state αh_{t-1} rather than at the current one. Writing this path-parametrised score as

$$S^{(\alpha)}_{j\to i} = h_{t-1}[j]\, \frac{\partial F_i}{\partial h_j}\bigg|_{\alpha h_{t-1}} = \phi'\big(a^{(\alpha)}[i]\big)\, W_{hh}[i,j]\, h_{t-1}[j],$$

our score is recovered as its value at the endpoint $\alpha = 1$,

$$S_{j\to i}(t) = S^{(1)}_{j\to i} = g_i(t)\, W_{hh}[i,j]\, h_{t-1}[j] \qquad (a^{(1)} = a_t,\ g_i(t) = \phi'(a_t[i])),$$

and the attribution is the average of $S^{(\alpha)}_{j\to i}$ over the path,

$$\mathrm{IG}_{j\to i} = \int_0^1 S^{(\alpha)}_{j\to i}\, d\alpha = \mathbb{E}_{\alpha\sim\mathcal{U}[0,1]}\big[S^{(\alpha)}_{j\to i}\big].$$

Because the wire and signal factors are constant in α, the two agree exactly iff the gate equals its own path average, $\phi'(a_t[i]) = \int_0^1 \phi'(a^{(\alpha)}[i])\, d\alpha$ — guaranteed when ϕ is affine along the segment $[\,0,\, a_t\,]$, so that the gate is constant in α (e.g. a ReLU unit that does not cross its threshold along the path), and approximate in the near-linear regime of a saturating nonlinearity. Otherwise, $S_{j\to i}(t)$ is the leading term of $\mathrm{IG}_{j\to i}$.

This is made clear by defining the recurrent drive as $r_i = (W_{hh}h_{t-1})[i]$ and rewriting

$$a^{(\alpha)}[i] = a_t[i] - (1-\alpha)\, r_i$$

so that a Taylor expansion of the gate about the endpoint (for ϕ twice differentiable along the path) gives

$$\phi'\big(a^{(\alpha)}[i]\big) = \phi'(a_t[i]) + (\alpha - 1)\, r_i\, \phi''(a_t[i]) + O(r_i^2),$$

and integrating over α (with $\int_0^1 (\alpha-1)\, d\alpha = -\frac{1}{2}$) yields

$$\mathrm{IG}_{j\to i} = S_{j\to i}(t) - \tfrac{1}{2}\, W_{hh}[i,j]\, h_{t-1}[j]\, r_i\, \phi''(a_t[i]) + O(r_i^2).$$

The zeroth-order term is exactly $S_{j\to i}(t)$, and the first correction is the postsynaptic curvature $\phi''(a_t[i])$ weighted by the recurrent drive r_i – vanishing when ϕ is affine ($\phi'' = 0$) or the recurrent contribution to the pre-activation is negligible.

B. Synthetic Benchmark Validation: Summary

Table 1 summarizes circuit discovery performance across the two synthetic benchmarks. Detailed per-task breakdowns, including per-position F1 scores for the ring attractor and temporal specificity, state-specific encoding, and noise robustness analyses for the HMM task, are provided in Appendices C and D respectively.

Table 1. Circuit discovery validation on synthetic benchmarks.

Task	**Network**	**Discovered (top-k)**	**Key Metric**
Ring Attractor	24 neurons	5 per position	83.67% mean F1
HMM Inference	128 neurons	5 per state	100% acc., 8.5× fewer params

C. Ring Attractor: Additional Details

We construct a 24-neuron network (19 excitatory, 5 inhibitory) implementing a ring attractor for spatial position encoding. Four positions are encoded through localized excitatory clusters: position 0 via neurons $\{0, 1, 2, 17, 18\}$, position 1 via neurons $\{2, 3, 4, 5, 6\}$, position 2 via neurons $\{7, 8, 9, 10, 11\}$, and position 3 via neurons $\{12, 13, 14, 15, 16\}$. We generate three trajectory types to test circuit discovery under different dynamics: (1) *static bumps* that maintain fixed positions for 20 timesteps, (2) *traveling bumps* that continuously drift from one position to another, and (3) *jumping bumps* with discrete transitions between non-adjacent positions.

Table 2 shows F1 scores of circuit discovery across trajectory types. For static trajectories, windowed causal interventions (top-k=5 neurons) consistently recover the ground truth position-encoding clusters, with F1 scores ranging from 80% to 100% across the four positions. The variation reflects the relative difficulty of disambiguating overlapping neuron assignments at position boundaries (e.g., neuron 2 participates in both position 0 and position 1 clusters).

For traveling bumps, F1 drops to 68% ± 35%, reflecting the dynamic nature of the computation: as the bump moves, different neurons become transiently important, and the ground truth itself becomes ambiguous across varying window. For jumping bumps (discrete position $0 \rightarrow 3 \rightarrow 1$ transitions), our method achieves 92% ± 10% F1. Critically, time-resolved analysis reveals dynamic reorganization during jumps: before the transition at t=8, neurons encoding the initial position (0) dominate causal importance; after the transition, neurons encoding the target position (3) become critical.

Table 2. Ring attractor circuit discovery performance.

Trajectory Type	**k neurons**	**F1 Score**	**Key Finding**
Static (pos 0)	5	80% ± 9%	Stable circuits
Static (pos 1)	5	85% ± 9%	Stable circuits
Static (pos 2)	5	100% ± 0%	Perfect recovery
Static (pos 3)	5	100% ± 0%	Perfect recovery
Average (static): **91% ± 10%** across positions			
Traveling bump	5	68% ± 35%	Dynamic reorganization
Jumping bump	5	92% ± 10%	Temporal dissociation

F1 score measures overlap between discovered top-k neurons and ground truth position-encoding clusters. Values show mean ± standard deviation across temporal windows within each trajectory type.

Beyond the quantitative metrics above, our time-resolved circuit visualizations provide direct insight into how the operative subcircuit reconfigures (or remains stable) across the three trajectory types.

Figure 4 shows the discovered circuit for position 2 at three widely spaced timesteps (t=5, 10, 15). The circuit comprises the expected position-encoding neurons $\{7, 8, 9, 10, 11\}$ together with inhibitory neurons 19 and 20. Node sizes and connectivity patterns are nearly identical across all three panels, with neurons 9 and 8 serving as dominant excitatory hubs throughout. Inhibitory nodes are present but contribute thin edges, consistent with a background modulatory role. This stability confirms that our pipeline correctly identifies a fixed maintenance circuit when the underlying computation is stationary.

Figure 5 captures a discrete position $0 \rightarrow 3$ transition around t=8. At t=7, the circuit is anchored by position-0 neurons (0, 1, 17, 18) with strong excitatory connectivity, while position-3 neurons (12–16) appear with small node sizes at the

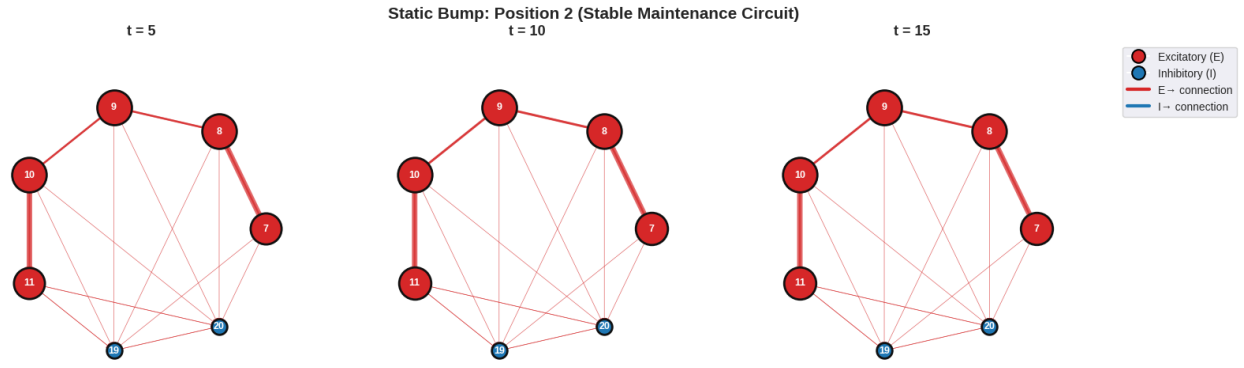

Figure 4. **Static bump: stable maintenance circuit for position 2.** Time-resolved circuit at t=5, 10, 15. The discovered circuit comprises position-2 neurons $\{7, 8, 9, 10, 11\}$ and inhibitory neurons $\{19, 20\}$. Node sizes and edge structure remain nearly identical across all three timesteps, confirming that the pipeline recovers a stable circuit when the underlying computation is stationary. Node color: excitatory (red), inhibitory (blue). Edge color: excitatory (red) vs. inhibitory (blue) influence. Edge width: magnitude of instantaneous influence $|S_{j\to i}(t)|$.

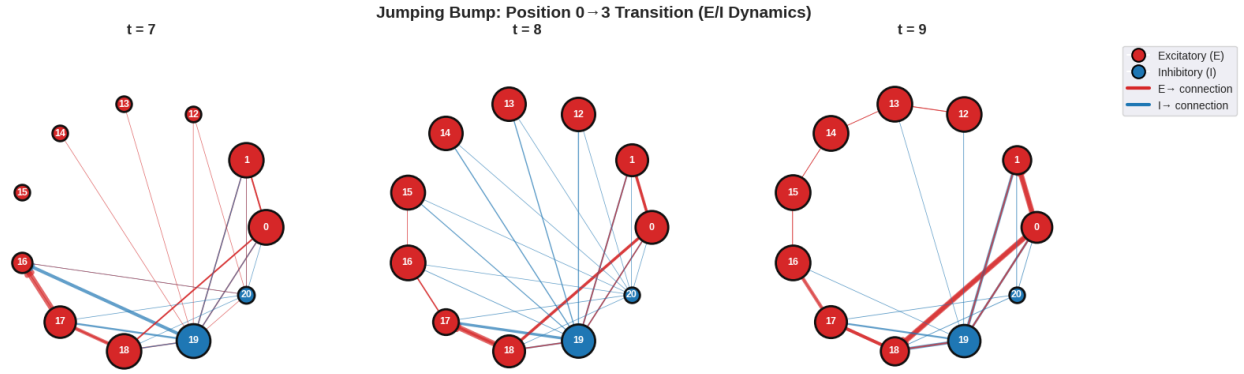

Figure 5. **Jumping bump: discrete position 0 → 3 transition.** Time-resolved circuit at t=7, 8, 9 around the jump. At t=7, position-0 neurons (0, 1, 17, 18) anchor the circuit. At the transition (t=8), position-3 neurons (14, 15, 16) grow while inhibitory node 19 strengthens its connections to both populations. By t=9, position-3 neurons (12, 13, 14) dominate with strong mutual excitation. The transient inhibitory involvement during the handoff parallels the gating motif observed in the flip-flop cue-switching circuit (Sec. 3.2, Fig. 2B).

periphery. At the transition (t=8), the circuit enters a mixed regime: position-3 neurons (14, 15, 16) grow in size and begin forming excitatory connections, while inhibitory node 19 thickens its edges toward both populations, mediating the handoff. By t=9, the reorganization is complete—position-3 neurons (12, 13, 14) dominate the circuit with strong mutual excitation, and position-0 neurons have receded. The transient recruitment of inhibitory drive during the transition parallels the gating motif observed in the flip-flop task (Sec. 3.2).

Figure 6 illustrates the gradual drift from position 0 to position 2, sampled at t=5, 12, and 20. At t=5, position-0 neurons (0, 1) are the largest excitatory nodes, with inhibitory neuron 19 providing the dominant non-excitatory influence. By t=12, the bump has reached position 1: neurons 4 and 5 have grown substantially, while nodes 0 and 1 begin to shrink; inhibitory neurons 19 and 20 both show strengthened edges, consistent with active redistribution of the bump. At t=20, the transition is complete—position-2 neurons (9, 10) and the intervening node 4 are now the largest, while position-0 neurons have become peripheral. Inhibitory node 20 has grown to be the largest node in the circuit, suggesting a sustained role in shaping the bump's final position. The smooth, progressive handoff of activity across the three snapshots visually confirms the continuous nature of the traveling bump dynamics and illustrates why F1 scores are lower for this trajectory type: the "correct" circuit is itself in flux at any given moment.

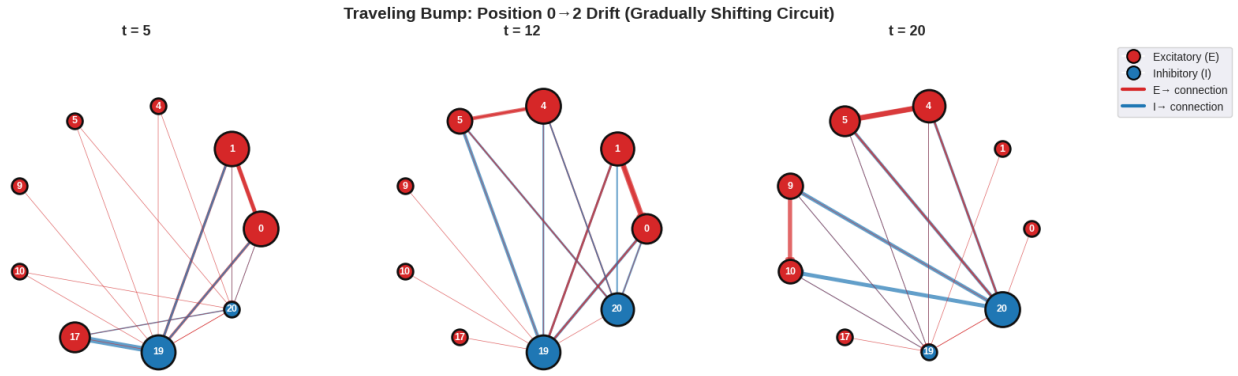

Figure 6. **Traveling bump: gradual position 0 → 2 drift.** Time-resolved circuit at t=5, 12, 20. Activity progressively transfers from position-0 neurons (0, 1) through position-1 neurons (4, 5) to position-2 neurons (9, 10). Inhibitory node 20 grows substantially over the course of the drift, suggesting a sustained role in redistributing the bump. The smooth handoff illustrates why F1 is lower for this trajectory type: the ground truth circuit is itself continuously shifting.

D. HMM Inference: Additional Details

We train a 128-neuron RNN to predict the next observation in sequences generated from a 3-state HMM with diagonal-dominant transition dynamics and state-specific emission patterns. The network implicitly learns to infer hidden states from observation history. Ground truth state beliefs are computed via Baum-Welch forward inference, providing oracle labels for validation.

Our circuit discovery method identifies state-specific maintenance circuits by applying windowed ablation during sequences that remain in a single state for extended periods. For each of the three states (A, B, C), we select the top-5 neurons by causal importance, yielding 15 neurons total. These discovered circuits achieve functional sufficiency: per-state decoder accuracy ranges from 96–99%, and combined 3-way classification reaches 100%—exactly matching the full 128-neuron baseline. This represents an 8.5× parameter reduction (15 vs. 128 neurons) with zero performance loss.

Temporal specificity. Figure 7 demonstrates that discovered neurons exhibit sharp temporal specificity: during state maintenance windows (shaded regions), discovered neurons (orange) show high causal importance as measured by ablation margin loss, while Baum-Welch correlation-selected neurons (green) show weaker and less temporally-localized importance. Critically, outside maintenance windows, discovered neuron importance drops to near-zero, confirming they encode state-specific computation rather than generic task features.

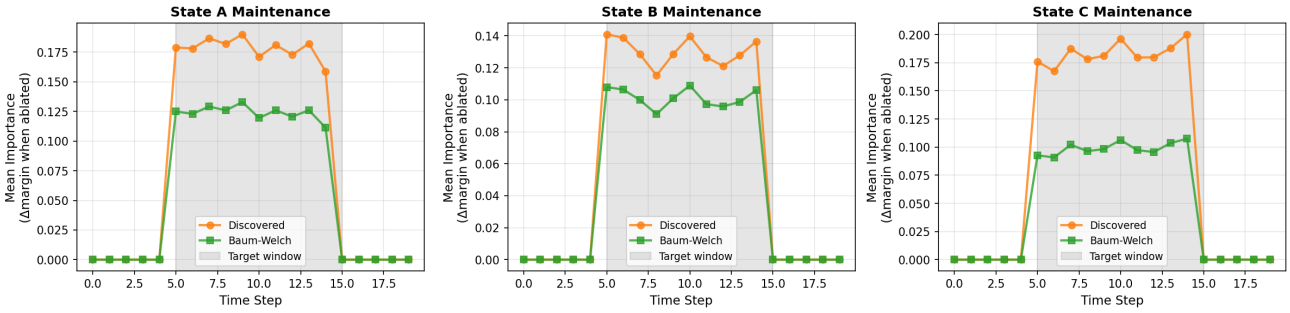

Figure 7. **Temporal specificity of discovered state circuits.** Mean causal importance (ablation-induced margin loss, Δ_{margin} when neuron ablated) over time for discovered (orange) vs. Baum-Welch correlation-based (green) neurons during maintenance of States A, B, and C. Target maintenance windows are shaded. Discovered neurons show high importance precisely during their state's maintenance window and near-zero elsewhere, demonstrating functional specificity. Baum-Welch neurons show weaker temporal localization.

State-specific encoding validation. To verify that discovered circuits encode their target states specifically, we test each state's 5-neuron circuit on a binary classification task: distinguishing "in state X" vs. "not in state X" (Fig. 8).

Discovered neurons achieve 95.3% average binary classification accuracy across the three states, significantly outperforming random neuron selection (92.6%) despite using only 5 neurons per state. Baum-Welch correlation-based selection achieves comparable performance (97.4%), indicating that both causal interventions and correlation-based methods successfully identify state-encoding neurons. Critically, all three states show consistent performance: discovered circuits achieve 93–97% accuracy per state, with State A and B neurons reaching near-perfect discrimination. This per-state consistency, combined with superior performance over random controls, confirms that windowed causal interventions identify functionally specialized circuits that systematically encode state-specific information rather than generic task-relevant features.

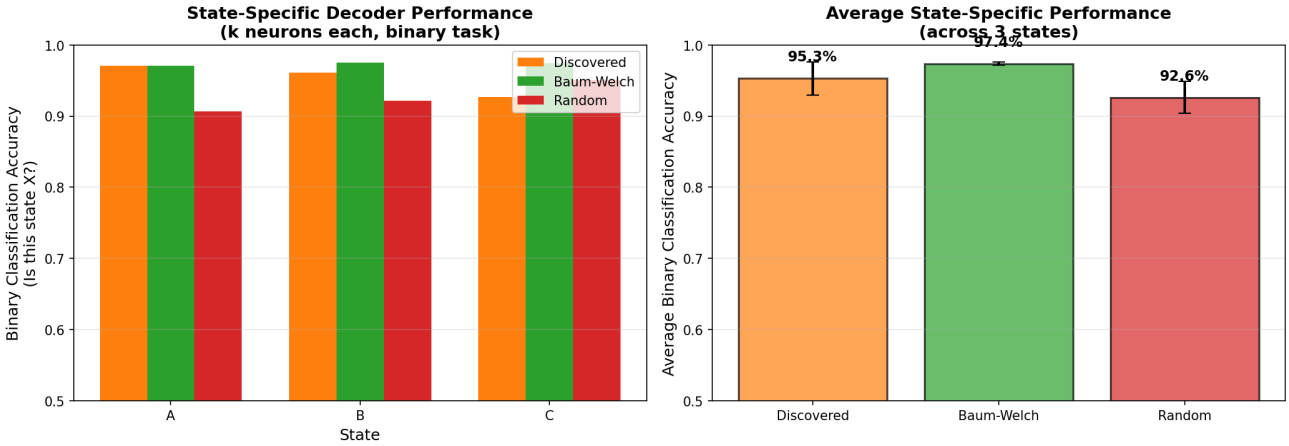

Figure 8. **State-specific encoding validation.** (Left) Binary classification accuracy ("Is this state X?") for state-specific neuron sets (5 neurons each). Discovered (orange), Baum-Welch correlation-based (green), and random (red) selections tested per state. (Right) Average performance across three states with error bars. Both discovered and Baum-Welch methods significantly outperform random controls, confirming systematic state-specific encoding.

Noise robustness. To test whether discovered circuits capture robust computational structure rather than mere correlations, we evaluate performance under additive Gaussian noise ($\sigma \in [0, 0.5]$) added to hidden states during decoding (Table 3). Across all noise levels, discovered circuits maintain an average accuracy of 77.7%, significantly outperforming correlation-based Baum-Welch neuron selection (68.3%). At the highest noise level (σ=0.5), discovered circuits retain 60% accuracy versus 55% for Baum-Welch selection, maintaining a consistent advantage. This demonstrates that windowed causal interventions identify functionally essential circuits that generalize under perturbation, whereas correlation-based methods conflate causal structure with spurious statistical dependencies.

Table 3. Noise robustness: per-state binary classification under additive Gaussian noise.

Method	**Mean Decoder Accuracy at Noise Level**					
	σ=0.0	σ=0.1	σ=0.2	σ=0.3	σ=0.4	σ=0.5
Discovered (5 per state)	97%	90%	81%	72%	66%	60%
Baum-Welch (5 per state)	90%	74%	68%	64%	59%	55%
Average across noise levels: **Discovered 77.7%** vs. Baum-Welch 68.3%						

Mean binary classification accuracy (5 neurons per state) averaged across three states (A, B, C) under additive Gaussian noise ($\mathcal{N}(0, \sigma^2)$). Our causally discovered circuits maintain consistent advantage across all noise levels compared to those identified using Baum-Welch.

E. Context-Dependent Flip-Flop Task: Diagnostic Stimuli

Figure 9 illustrates the general task structure and the three diagnostic stimulus families used to isolate specific computational demands in the context-dependent flip-flop task.

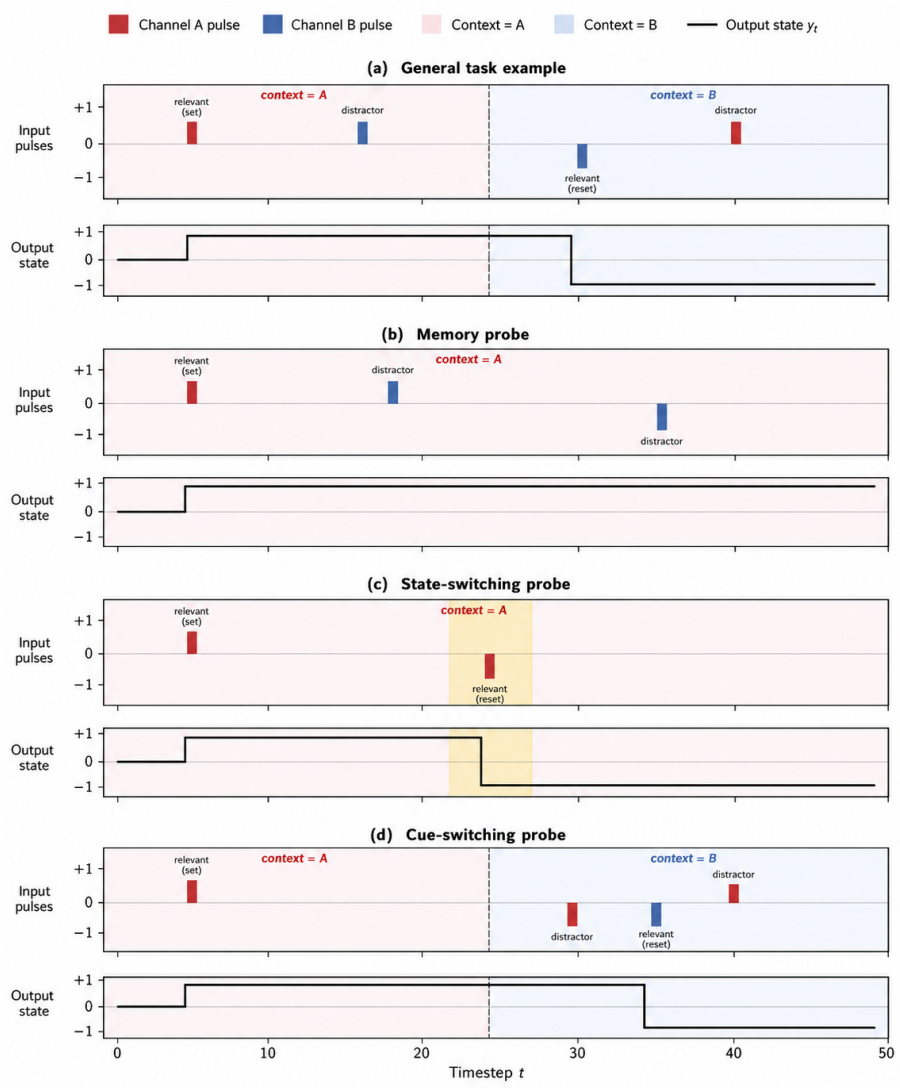

Figure 9. **Context-dependent flip-flop task and diagnostic stimuli.** (a) General task structure: the network receives pulses on two channels (red: A, blue: B) and a context signal (background shading) determining which channel is currently relevant. The output state y_t (black) updates only on relevant pulses and persists otherwise; pulses on the irrelevant channel are distractors. (b) Memory probe: a single relevant pulse initializes the state, followed by a long delay with only irrelevant distractors; correct behavior requires stable maintenance of the set state. (c) State-switching probe: a relevant reset pulse (yellow: analysis window) flips the state under a fixed context with no distractors, isolating the transient switching machinery. (d) Cue-switching probe: after establishing a state under context A, the context flips to B; subsequent distractor pulses on the now-irrelevant channel A must be ignored, while a relevant pulse on channel B drives the state update.

F. Flip-Flop RNN: Model Architecture and Training Details

Task and targets. Sequences have length T=20 and three input channels per step: two pulse channels $(x_t^{(1)}, x_t^{(2)})$ and a binary context $c_t \in \{-1, +1\}$ indicating which pulse is relevant. Training sequences are sampled i.i.d.; pulses on the two channels are drawn with low per-step probability and amplitudes ± 1; the context c_t flips with small probability (≈ 0.05) to elicit cue-switch trials. The latent flip-flop state $s_t \in \{-1, 0, +1\}$ is initialized at 0 and updates only on relevant pulses (per the current context); otherwise it persists. For cross-entropy, the target at each step is the class index

$$y_t^{\text{CE}} = (s_t + 1) \in \{0, 1, 2\} \quad (-1 \mapsto 0,\ 0 \mapsto 1,\ +1 \mapsto 2).$$

By default, steps before the first relevant pulse are supervised as the 0-class.

Network. We use a single-layer, Dale-constrained RNN with a $\tanh$-then-rectify hidden nonlinearity constructed and trained as per (Balwani et al., 2025b):

$$\begin{aligned}
a_t &= W_{ih}\, x_t + W_{hh}\, h_{t-1} + b_h, \\
u_t &= \tanh(a_t), \\
h_t &= [u_t]_+ = \max(u_t, 0) \in \mathbb{R}^N, \\
z_t &= W_o\, h_t + b_o \in \mathbb{R}^3, \qquad \hat{y}_t = \mathrm{softmax}(z_t).
\end{aligned}$$

Hidden units are partitioned into excitatory $\mathcal{E}$ and inhibitory $\mathcal{I}$ populations in the ratio $\mathcal{E} : \mathcal{I} = 8 : 2$ with 128 hidden neurons in total. Dale's law is enforced on the recurrent matrix W_{hh}; input W_{ih} and readout W_o are unconstrained.

Training objective. We minimize stepwise cross-entropy over the three classes:

$$\mathcal{L} = \frac{1}{BT} \sum_{b=1}^{B} \sum_{t=1}^{T} \mathrm{CE}\big(z_t^{(b)},\, y_t^{\text{CE},(b)}\big),$$

where B is the batch size and CE is the standard logits-based cross-entropy loss.

Optimization and Dale projection. Parameters are updated with Adam (learning rate 10^{-3}; other hyperparameters at PyTorch defaults) for 200 epochs with batch size 200. After each optimizer step, the recurrent matrix is projected onto the sign-constrained set:

$$(W_{hh})_{ij} \leftarrow \begin{cases} \max\{(W_{hh})_{ij},\, 0\}, & \text{if presynaptic } j \in \mathcal{E}, \\ \min\{(W_{hh})_{ij},\, 0\}, & \text{if presynaptic } j \in \mathcal{I}. \end{cases}$$

This is the closest sign-consistent update in the Frobenius sense and maintains Dale's law throughout training.

Nonlinearity and postsynaptic gain. With $h_t = [\tanh(a_t)]_+$, the postsynaptic gain used in our time-resolved linearization (Sec. 2.2) takes the specific form

$$D_{\phi,t} = \mathrm{diag}\big((1 - u_t^2) \odot \mathbf{1}\{h_t > 0\}\big), \quad g_i(t) = D_{\phi,t}(i,i) \geq 0.$$

Under Dale's law, the sign of an instantaneous influence $S_{j \to i}(t) = g_i(t)\, W_{hh}[i,j]\, h_{t-1}[j]$ matches the synaptic sign, simplifying E/I interpretation in our circuit visualizations.

G. State-Switching Circuit Visualization

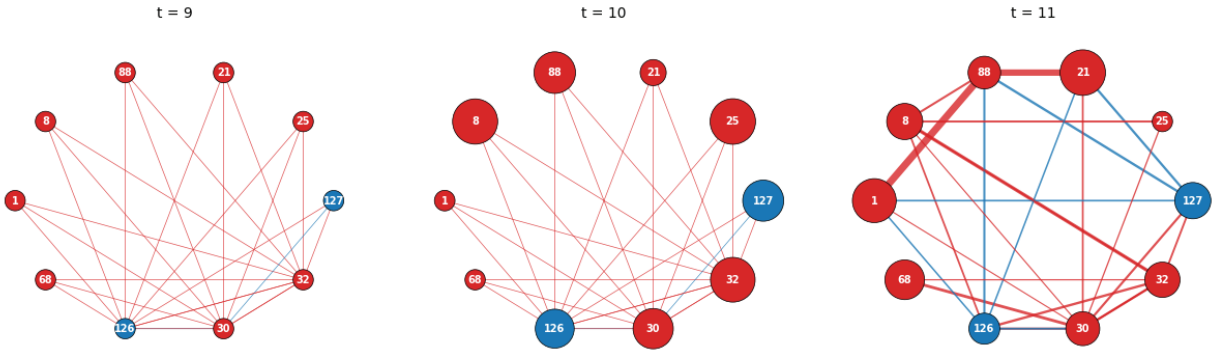

Figure 10. **Time-resolved switching circuit.** Switching circuit at timesteps t=9, 10, 11 around the switching event at t=10. Only the top-10 neurons (by windowed causal score) are shown. Node color: excitatory (red), inhibitory (blue). Node size: proportional to normalized hidden activity at time t. Edge color: sign of effective influence. Edge width: magnitude of instantaneous influence $|S_{j\to i}(t)|$. Compared to the cue-switching circuit (Fig. 2B), the inhibitory dynamics here are weaker and slightly delayed, consistent with the need to overcome the stable excitatory memory loop to drive the state flip.

H. V1 RNN: Dataset and Stimuli

The Allen Institute Visual Behavior 2P dataset (Garrett et al., 2023; Allen Institute for Brain Science, 2021) provides two-photon calcium recordings from mouse visual cortex during the change-detection task, with images drawn from a familiar set seen during training and a novel set introduced only at test; on $\sim$5% of trials the second image of a presentation is omitted and replaced by a blank screen. We model the half-set window beginning at this second presentation (or its omission) – the 250 ms presentation plus the following 500 ms interval, binned into 6 steps of 125 ms (750 ms total). Training targets are the trial-averaged $\Delta F/F$ responses of each V1 L2/3 cell-type population, curated as in Balwani et al. (2025b).

I. V1 RNN: Model Architecture and Training Details

The model follows the CelltypeRNN architecture of Balwani et al. (2025b) (Sec. 4), restricted to Layer 2/3 of mouse primary visual cortex. It contains 240 recurrent neurons partitioned into three biologically identified populations: 192 pyramidal (Pyr, indices 0–191, excitatory), 32 SST (indices 192–223, inhibitory), and 16 VIP (indices 224–239, inhibitory). The hidden nonlinearity is $\tanh$ followed by a rectifying clamp, as in the flip-flop model (Appendix F). Dale's law is enforced on W_{hh}: Pyr neurons have non-negative outgoing weights, while SST and VIP neurons have non-positive outgoing weights. Inter-population connectivity follows the sparse wiring patterns measured by Campagnola et al. (2022), and each cell type receives input only from its corresponding channel via cell-type-specific input masking on W_{ih}.

Table 4. **Intra-laminar connection probabilities among the V1 L2/3 cell-type populations.** Taken from the measurements of Campagnola et al. (2022) and restricted to L2/3 (cf. their full multi-layer table). Columns index the presynaptic (source) population, rows the postsynaptic (target) population; each entry is the probability of a connection for the corresponding source–target pair. These probabilities define the sparse inter-population connectivity targets enforced via top-prob pruning.

	Source (presynaptic)		
Target (postsynaptic)	Pyr	SST	VIP
Pyr	0.06	0.23	0.05
SST	0.30	0.05	0.14
VIP	0.16	0.30	0.01

Following Balwani et al. (2025b), the network is constructed by first training a dense Elman RNN with Dale's backpropagation, and subsequently sparsifying it to the cell-type connectivity targets of Campagnola et al. (2022) via top-prob pruning (Table 4), and fine-tuning the pruned network with Dale's backpropagation at fixed sparsity. The Pyr:SST:VIP ratio is 12:2:1 (excitatory:inhibitory $= 4$:1), and the per-population linear readout is randomly initialized and held fixed during training. We optimize with Adam (learning rate 10^{-3}, PyTorch defaults otherwise), training the dense model for 100 epochs and fine-tuning for 50.

Since the network is trained to predict neural activity rather than to classify conditions, we define the decision margin for our causal interventions by training per-timestep logistic regression classifiers (scikit-learn `LogisticRegression`) on the hidden state population to discriminate Familiar from Omission trials. The margin at timestep t is the signed decision function $m_t = w_t^\top h_t + b_t$, oriented so that positive values indicate correct classification. Classifiers are fit independently at each of the 6 timesteps on training data and evaluated on a held-out set of 1000 trials (500 per condition).

J. V1 Analysis: Cell-Type Ablation Profiles

Figure 11 shows the per-timestep ablation importance for each cell type under both conditions. Pyramidal neurons (left) show modest, condition-independent importance with wide confidence intervals, reflecting their large population size (192 neurons) and distributed role. SST neurons (center) show elevated importance for familiar trials at intermediate timesteps, consistent with a role in processing expected stimuli. VIP neurons (right) show the striking crossover pattern discussed in the main text (Sec. 4): high importance for familiar trials early, switching to high importance for omission trials late, with the inversion occurring around t=2–3.

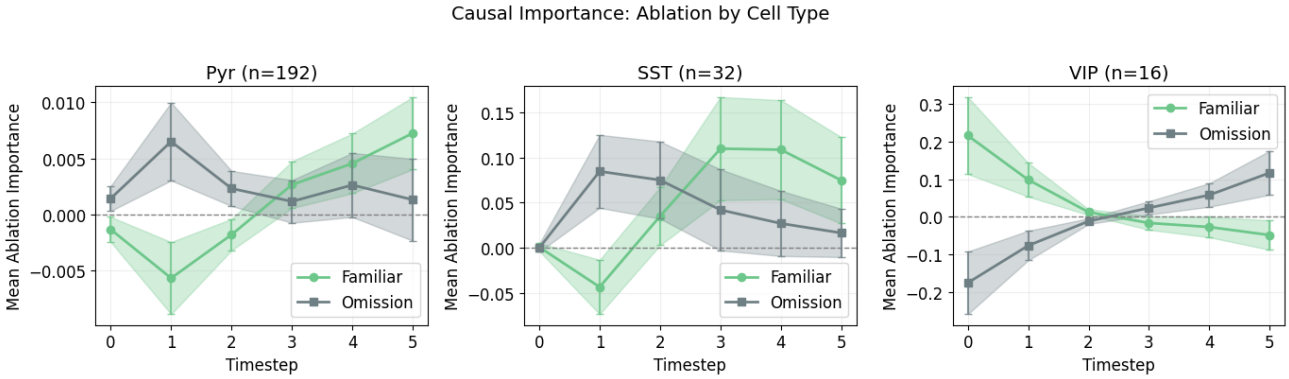

Figure 11. **Causal importance by cell type and condition.** Mean ablation importance (margin drop upon cell-type ablation) for Pyr (n=192), SST (n=32), and VIP (n=16) populations across 6 timesteps, shown separately for Familiar (green) and Omission (grey) conditions. Shaded regions indicate 95% confidence intervals across trials. VIP neurons exhibit a pronounced temporal crossover absent in the other populations.

K. V1 Analysis: Hidden State Dynamics by Cell Type

Figure 12 shows the mean hidden state activity for each cell type under both conditions, alongside the condition difference (Omission − Familiar). VIP neurons show a characteristic ramping pattern: activity decreases initially in both conditions but then diverges, with omission trials showing sustained ramping at late timesteps while familiar trials plateau. This ramping profile is consistent with the interval-timing hypothesis of (Huang et al., 2025), which posits that VIP activity reflects elapsed time since the last stimulus-driven reset. Notably, the condition difference for VIP (bottom right) transitions from negative (lower activity in omission) at early timesteps to strongly positive (higher activity in omission) at late timesteps, mirroring the causal crossover in ablation importance.

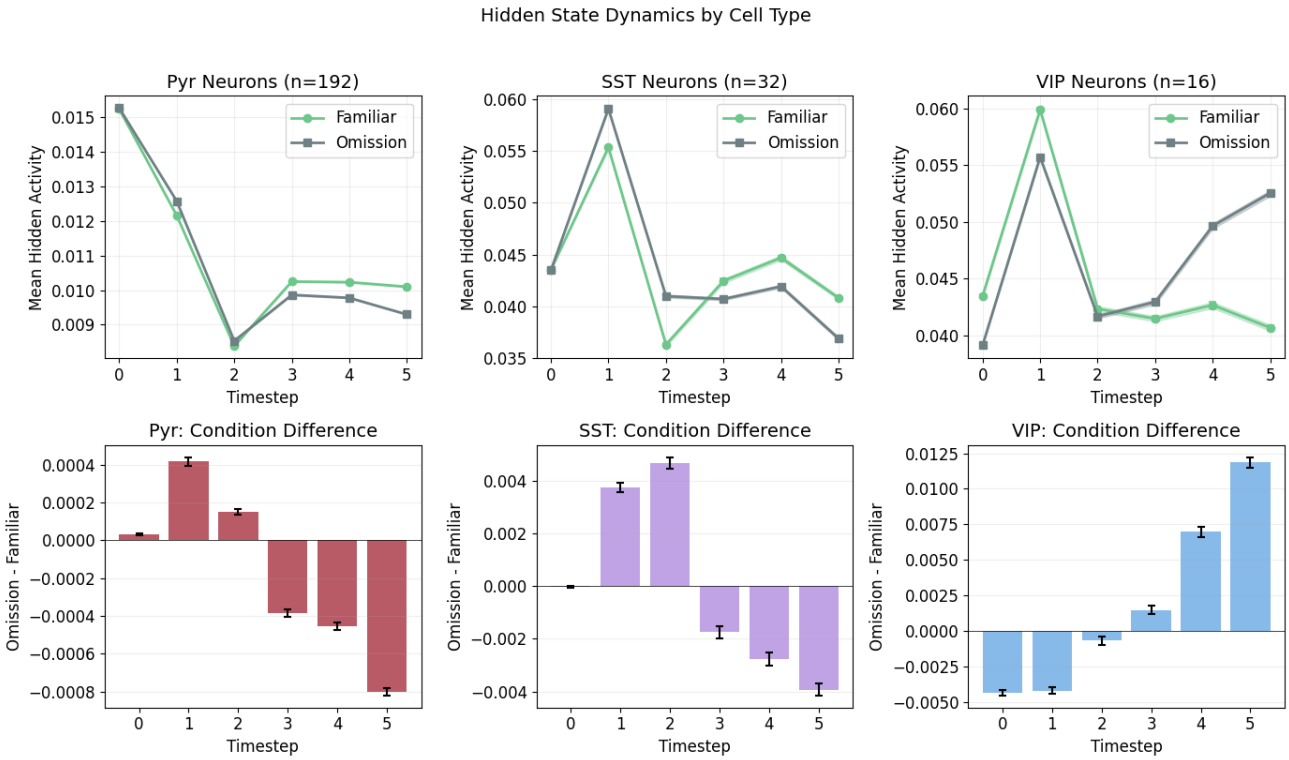

Figure 12. **Hidden state dynamics by cell type.** Top row: Mean hidden state activity for Pyr, SST, and VIP populations under Familiar (green) and Omission (grey) conditions. Bottom row: Condition difference (Omission − Familiar) with 95% confidence intervals. VIP neurons show a late-emerging positive difference consistent with ramping dynamics during stimulus omission.

L. V1 Analysis: Correlational vs. Causal Dissociation

Figure 13 compares three population-level measures of VIP's role in omission processing: correlational activity difference (Omission − Familiar), classifier weight magnitude, and causal ablation importance difference. While the correlational measures (panels A–B) suggest that VIP's role grows monotonically over time, the causal measure (panel C) reveals the non-monotonic crossover, demonstrating that activity difference and classifier weight are insufficient proxies for causal contribution. The bottom row isolates VIP, showing that all three measures eventually agree at late timesteps but diverge at early timesteps—precisely the regime where the causal intervention reveals VIP's initially interfering role.

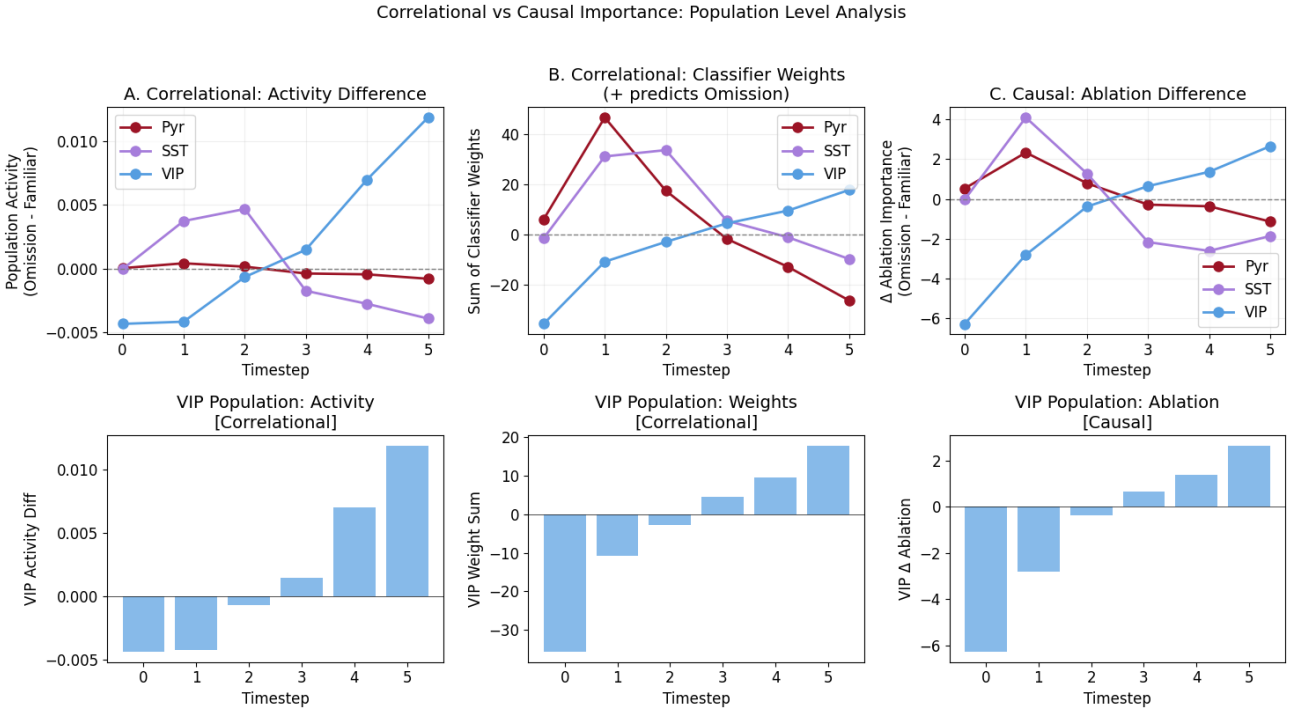

Figure 13. **Correlational vs. causal importance at the population level.** Top row: Population-level comparison across Pyr (red), SST (purple), and VIP (blue) for (A) activity difference, (B) classifier weights, and (C) causal ablation importance difference (Omission − Familiar). Bottom row: VIP-specific bar charts for each measure. The correlational measures suggest monotonically increasing VIP importance, while the causal measure reveals the early-phase inversion.

M. V1 Analysis: Complete Effective Connectivity

Figure 14 shows the signed effective connectivity for all six inter-population pathways computed via recurrence linearization. Inhibitory pathways (blue shading: VIP→SST, SST→Pyr, SST→VIP, VIP→Pyr) show uniformly negative connectivity, consistent with Dale's law constraints. Excitatory pathways (red shading: Pyr→SST, Pyr→VIP) show positive connectivity. Condition-dependent divergence is most pronounced in the VIP→SST and VIP→Pyr pathways at late timesteps, consistent with the causal crossover.

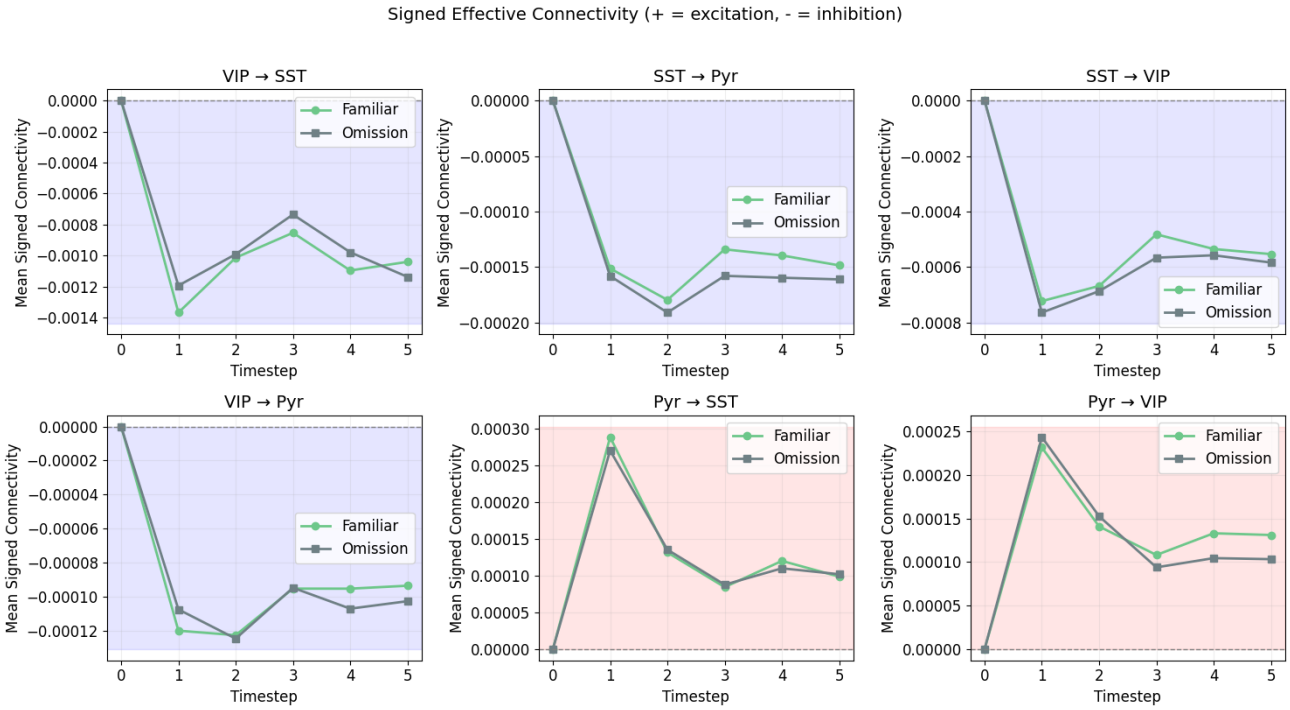

Figure 14. **Signed effective connectivity for all inter-population pathways.** Mean signed connectivity $S_{j\to i}(t)$ aggregated by source and target population, shown for Familiar (green) and Omission (grey) across 6 timesteps. Background shading indicates inhibitory (blue) vs. excitatory (red) pathways. All signs are consistent with Dale's law constraints.

N. V1 Analysis: Disinhibitory Pathway Details

Figure 15 provides a detailed view of the VIP→SST→Pyr disinhibitory pathway. Panels A–C show the two individual legs (VIP→SST, SST→Pyr) and their product (the two-hop disinhibitory pathway strength) for both conditions. Panels D–F show the condition difference (Omission − Familiar), confirming that the disinhibitory pathway is weaker for omission early but stronger late. Panel G reproduces VIP ablation importance for comparison, and panel H shows the trial-by-trial correspondence between disinhibitory pathway strength and VIP ablation effect, color-coded by timestep. The positive correlation at late timesteps (warm colors, upper-right quadrant) confirms that stronger disinhibitory engagement predicts greater VIP causal importance on a per-trial basis.

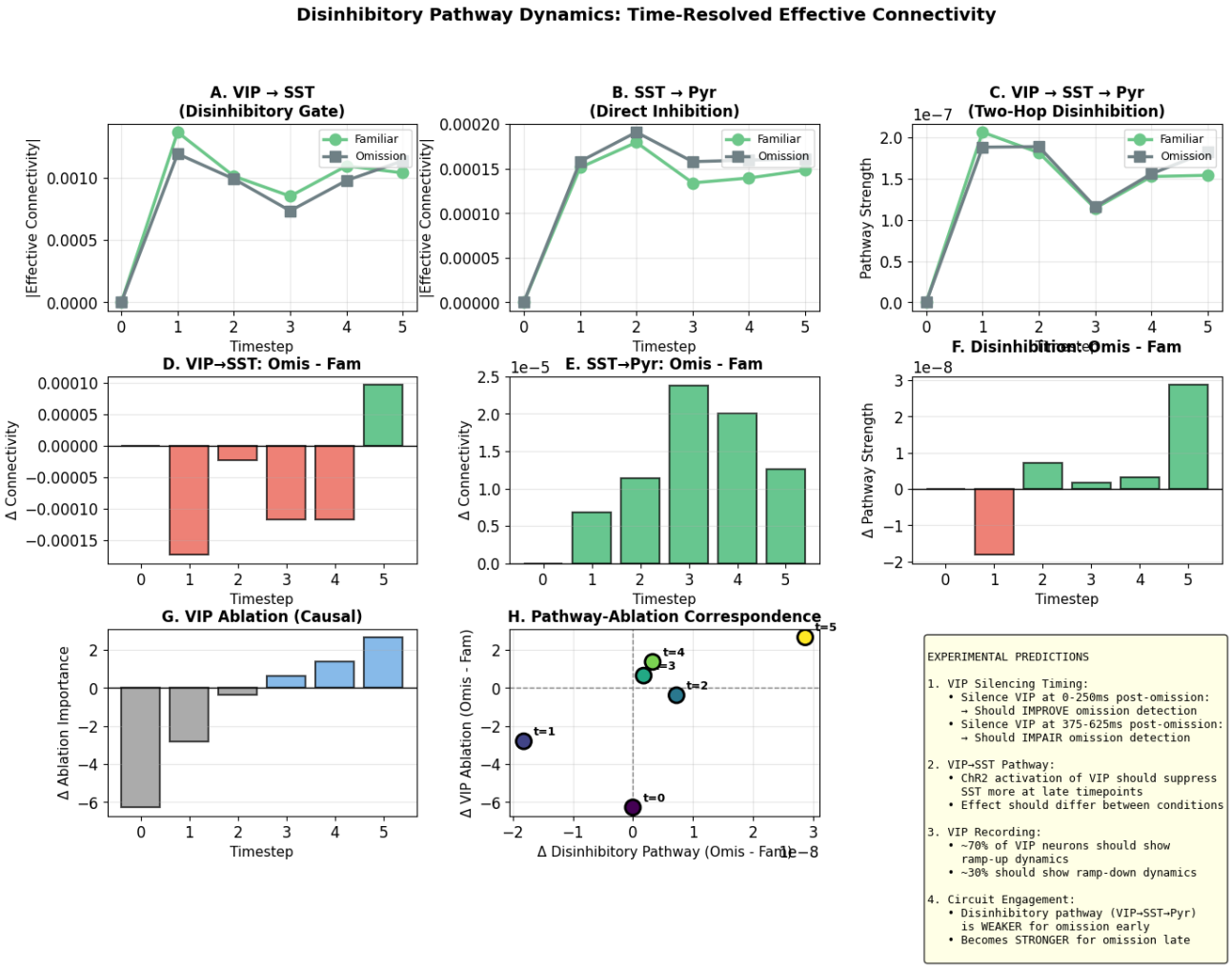

Figure 15. **Disinhibitory pathway dynamics: time-resolved effective connectivity.** Panels A–C: Absolute effective connectivity for VIP→SST (disinhibitory gate), SST→Pyr (direct inhibition), and the two-hop VIP→SST→Pyr pathway for Familiar (green) and Omission (grey). Panels D–F: Condition differences (Omission − Familiar). Panel G: VIP ablation importance for comparison. Panel H: Trial-by-trial scatter of disinhibitory pathway strength vs. VIP ablation effect, color-coded by timestep.

O. V1 Analysis: Interpreting the VIP Crossover

Here we expand on the mechanistic interpretation of the VIP causal crossover described in Sec. 4.

Consistency with predictive processing at late timesteps. At late timesteps (t=3–5), VIP ablation selectively impairs omission classification: VIP is indeed necessary for processing the unexpected stimulus absence, consistent with the dual-circuit hypothesis of (Najafi et al., 2025).

Early-phase inversion and the timing hypothesis. The early-phase pattern (t=0–2) is inconsistent with a pure prediction-error account, which would predict VIP necessity throughout the trial. Instead, VIP ablation *improves* omission classification early, indicating that VIP activity actively interferes with correct discrimination at these timesteps. This aligns with the interval-timing account of (Huang et al., 2025): early VIP activity reflects an intrinsic timing signal that has not yet been reinterpreted as surprise, and its presence during omission trials—where VIP has not been reset by a stimulus-driven transient—initially confuses the downstream readout. For familiar trials the pattern is reversed: VIP is causally important early (the timing signal carries "saw stimulus" information) but becomes dispensable late (VIP has recovered and is no longer informative).

The double crossover as evidence of functional mode transition. The trial-by-trial motif–ablation correlation (Fig. 3C) reveals that the relationship between the VIP→SST→Pyr disinhibitory motif strength and VIP's causal importance flips sign across both conditions and time. At early timesteps, the motif positively predicts importance on familiar trials (r=0.74) but negatively on omission trials (r=−0.58), indicating that the same circuit motif carries "correct" information under one condition and "wrong" information under the other. By late timesteps, this relationship inverts for both conditions. This double crossover indicates that VIP's functional mode—not merely its activity level—transitions over time, consistent with a reinterpretation of the timing signal by downstream circuits rather than a change in VIP's own computational role.

Synthesis. Taken together, these findings suggest that VIP's intrinsic timing dynamics provide the temporal substrate upon which prediction error signals are constructed: the ramping activity accumulates elapsed time, and after ∼300–400ms the downstream VIP→SST→Pyr circuit reinterprets this timing signal as evidence of stimulus absence. The ∼300–400ms reconfiguration timescale constitutes a concrete experimental prediction: temporally-targeted VIP silencing at early (<250ms) vs. late (>375ms) post-omission windows should produce opposite effects on omission detection, improving and impairing performance respectively.