Flexible gating between subspaces by a disinhibitory motif: a neural network model of internally guided task switching

Yue Liu*, Xiao-Jing Wang*
New York University

Abstract

Behavioral flexibility relies on the brain's ability to switch rapidly between multiple tasks, even when the task rule is not explicitly cued but must be inferred through trial and error. The underlying neural circuit mechanism remains poorly understood. We investigated recurrent neural networks (RNNs) trained to perform an analog of the classic Wisconsin Card Sorting Test. The networks consist of two modules responsible for rule representation and sensorimotor mapping, respectively, where each module is comprised of a circuit with excitatory neurons and three major types of inhibitory neurons. We found that rule representation by self-sustained persistent activity across trials, error monitoring and gated sensorimotor mapping emerged from training. Systematic dissection of trained RNNs revealed a detailed circuit mechanism. The networks' dynamical trajectories for different rules reside in separate subspaces of population activity; they become virtually identical and performance was reduced to chance level when dendrite-targeting somatostatin-expressing interneurons were silenced, demonstrating that rule-based gating critically depends on the disinhibitory motif.

Introduction

A signature of cognitive flexibility is the ability to adapt to a changing task demand. Oftentimes, the relevant task is not explicitly instructed, but need to be inferred from previous experiences. In laboratory studies, this behavioral flexibility is termed un-cued task switching. A classic task to evaluate this ability is the Wisconsin Card Sorting Test (WCST) [1]. During this task, subjects are presented with an array of cards, each with

*Corresponding author. Email: yl4317@nyu.edu, xjwang@nyu.edu

multiple features, and should respond based on the relevant feature dimension (i.e. the task rule) that changes across trials. Crucially, subjects are not instructed on when the rule changes, but must infer the currently relevant rule based on the outcome of previous trials. Intact performance on un-cued task switching depends on higher-order cortical areas such as the prefrontal cortex (PFC) [2, 3, 4, 5, 6], which has been proposed to represent the task rule and modulate the activity of other cortical areas along the sensorimotor pathway [7].

Four essential neural computations must be implemented by the neural circuitry underlying un-cued task switching. First, it should maintain an internal representation of the task rule across multiple trials when the rule is unchanged. Second, soon after the rule switches, the animal will inevitably make errors and receive negative feedback, since the switches are un-cued. This negative feedback should induce an update to the internal representation of the task rule. Third, the neural signal about the task rule should be communicated to the brain regions responsible for sensory processing and action selection. Fourth, this rule signal should be integrated with the incoming sensory stimulus to produce the correct action.

Prior work has identified neural correlates of cognitive variables presumed to underlie these computations including rule [8], feedback [8, 9, 10] and conjunctive codes for sensory, rule, and motor information [11]. In addition, different types of inhibitory neurons are known to play different functional roles in neural computation: while parvalbumin (PV)-expressing interneurons are suggested to underlie feedforward inhibition [12], interneurons that express somatostatin (SST) and vasoactive intestinal peptide (VIP) have been proposed to mediate top-down control [13, 14, 15, 16]. In particular, SST and VIP neurons form a disinhibitory motif [17, 18, 19] that has been hypothesized to instantiate a gating mechanism for flexible routing of information in the brain [20]. However, there is currently a lack of mechanistic understanding of how these neural representations and cell-type-specific mechanisms work together to accomplish un-cued task switching.

To this end, we used computational modeling to formalize and discover mechanistic hypotheses. In particular, we used tools from machine learning to train a collection of biologically informed recurrent neural networks (RNNs) to perform an analog of the WCST used in monkeys [21, 8, 9]. Training RNN [22] does not presume a particular circuit solution, enabling us to explore potential mechanisms. For this purpose, it is crucial that the model is biologically plausible. To that end, each RNN was set up to have two modules: a “PFC” module for rule maintenance and switching and a “sensorimotor” module for executing the sensorimotor transformation conditioned on the rule. To explore the potential functions of different neuronal types in this task, each module of our network consists of excitatory neurons with somatic and dendritic compartments as well as PV, SST and VIP inhibitory neurons, where the connectivity between cell types is constrained by experimental data (Methods).

After training, we performed extensive dissection of the trained models to reveal that close interplay between local and across-area processing was essential for solving the WCST. First, we found that abstract cognitive variables are distinctly represented in the PFC module. In particular, two subpopulations of excitatory neurons emerge in the PFC module - one encodes the task rule and the other shows nonlinear mixed-selectivity for rule and negative feedback. Notably, neurons with similar response profiles have been reported in neurophysiological recordings of monkeys performing the same task [8, 9]. Second, we identified interesting structures in the local connectivity between different neuronal assemblies within the PFC module, which enabled us to compress the high-dimensional PFC module down to a low-dimensional simplified network. Importantly, the neural mechanism for maintaining and switching rule representation is readily interpretatble in the simplified network. Third, we found that the rule information in the PFC module is communicated to the sensorimotor module via structured long-range connectivity patterns along the monosynaptic excitatory pathway, the di-synaptic pathway that involves PV neurons, as well as the trisynaptic disinhibitory pathway that involves SST and VIP neurons. In addition, different dendritic branches of the same excitatory neuron in the sensorimotor module can be differentially modulated by the task rule depending on the sparsity of the local connections from the dendrite-targeting SST inhibitory neurons. Fourth, single neurons in the sensorimotor module show nonlinear mixed selectivity to stimulus, rule and response, which crucially depends on the activity of the SST neurons. On the population level, the neural trajectories for the sensorimotor neurons during different task rules occupy nearly orthogonal subspaces, which is disrupted by silencing the SST neurons. Lastly, we found heterogeneous but structured patterns of input connections to the excitatory neurons in the sensorimotor module. This structure is predicted by a mathematical argument and enables appropriate rule-dependent action selection. These results are consistent across dozens of trained RNNs with different types of dendritic nonlinearities and initializations, therefore pointing to a common neural circuit mechanism underlying the WCST.

Results

Training modular recurrent neural networks with different types of inhibitory neurons

We trained a collection of modular RNNs to perform the WCST. Each RNN consists of two modules: the “PFC” module receives an input about the outcome of the previous trial, and was trained to output the current rule; the “sensorimotor” module receives the sensory input and was trained to generate the correct choice. Each module is endowed with excitatory neurons with somatic and two dendritic compartments, as well as three major types of genetically-defined inhibitory neurons (PV, SST and VIP). Different types of neu-

rons have different inward and outward connectivity patterns constrained by experimental data in a binary fashion (Methods, Figure 1b). The somata of all neurons are modeled as standard leaky units with a rectified linear activation function (Methods). The activation of the dendritic compartments, which can be viewed as a proxy for the dendritic voltage, is a nonlinear sigmoidal function of the excitatory and inhibitory inputs they receive (Methods). The specific form of the nonlinearity is inspired by experiments showing that inhibition acts subtractively or divisively on the dendritic nonlinearity function depending on its relative location to the excitation along the dendritic branch [23]. Therefore, we trained RNNs with both subtractive and divisive dendritic nonlinearities to explore their effects on the network function (Methods).

The task we trained the network on is a WCST variant used in monkey experiments [21, 8, 9] (Figure 1a). During each trial, a reference card with a particular color and shape is presented on the screen for 500 ms, after which three test cards appear around the reference card for another 500 ms. Each card can have one of the two colors (red or blue) and one of the two shapes (square or triangle). A choice should be made to indicate the test card that has the same relevant feature (color or shape) as the reference card, after which the outcome of the trial is given, followed by an inter-trial interval. The relevant feature to focus on, or the task rule, changes randomly every few trials. Critically, the rule changes were not cued, requiring the network to memorize the rule of the last trial using its own dynamics. Therefore, the network dynamics should be carried over between consecutive trials, rather than reset at the end of each trial as has been done traditionally [24, 25]. To this end, the network operated continuously across multiple trials during training, and the loss function was aggregated across multiple trials (Methods). We use supervised learning to adjust the strength of all the connections (input, recurrent and output) by minimizing the mean squared error between the output of both modules and the desired outputs. Notably, only the connections between certain cell types are non-zero and can be modified. This is achieved using a mask matrix, similar to [26] (Methods).

After training converged, we tested the models by running them continuously across 100 trials of WCST with 10 rule switches at randomly chosen trials. The networks were able to recover their performance after experiencing only a single error trial in the new rule (Figure 1c, d). Single neurons from both modules exhibit rule-modulated activity (Supplementary Figure 1). In the following sections, we will “open the black box” to understand the mechanism the networks used to perform the WCST.

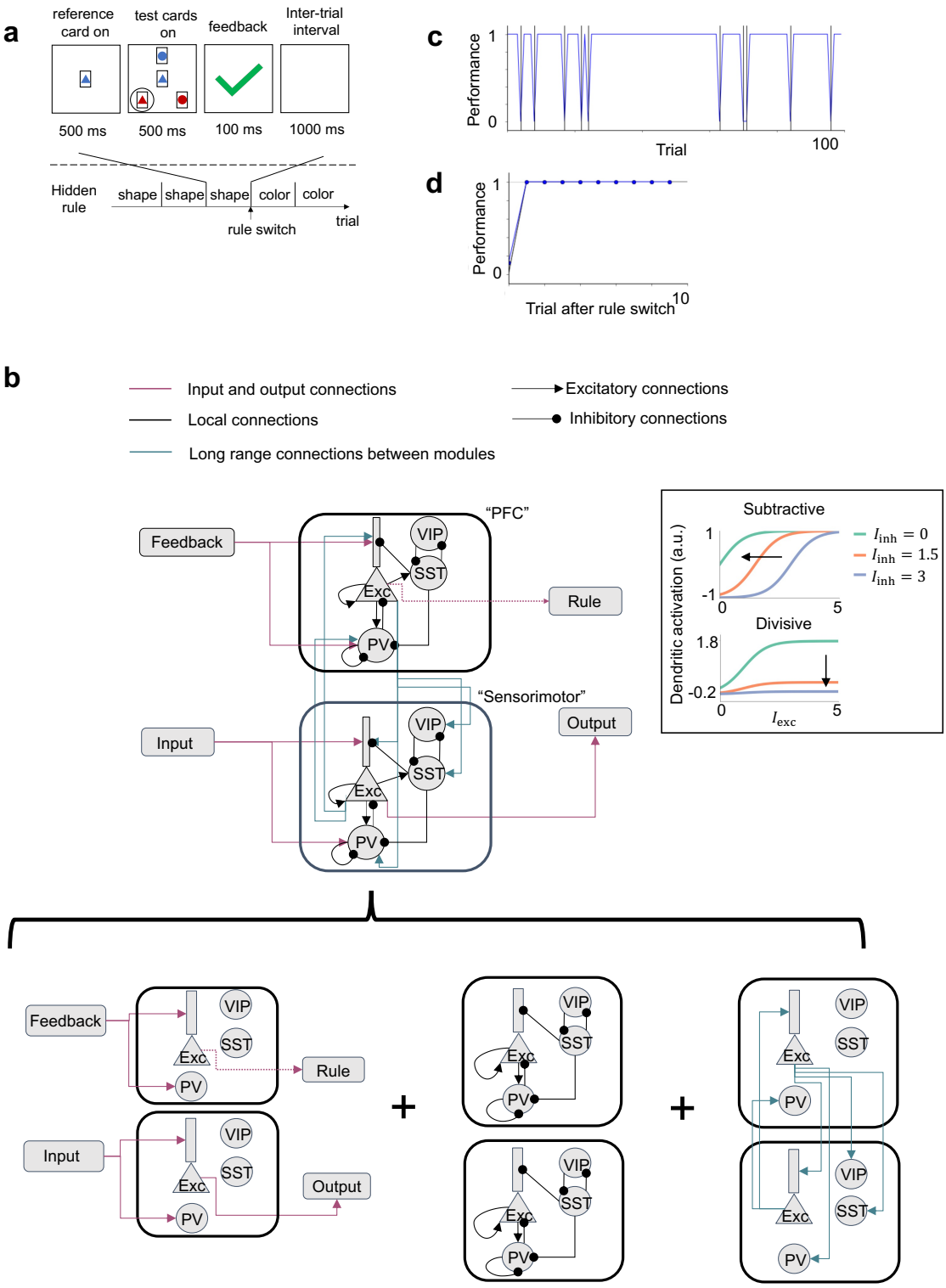


Figure 1: (Caption next page.)

Figure 1: (Previous page.) **Model setup and performance.**
a. The schematic of the WCST task. Subjects are required to choose the card that matches the reference card at the center according to a hidden rule that switches after a number of trials.
b. The RNN contains a "PFC" module and a "sensorimotor" module. The PFC module receives an input about the feedback of the previous trial, and is trained to produce the current rule. The sensorimotor module receives the sensory input and is trained to produce the correct choice. Each module is endowed with excitatory neurons and three types of inhibitory neurons: PV, SST and VIP. The cell-type-specific connectivity is constrained by experimental data (Methods). Bottom part shows the decomposition of the model architecture into the input and output target (left, magenta. The dashed line from PFC to rule represents the fact that the PFC module was trained to represent the rule but there are no explicit rule output nodes in the model), the local connectivity (middle, black) and inter-modular connections (green, right). All connections were trained. Each excitatory neuron is modeled with a somatic and two dendritic compartments. Inset shows for the two types of dendritic nonlinearities used the relationship between the excitatory input onto the dendrite and the dendritic activity for different levels of inhibitory inputs.
c. The performance of the model during testing, for an example network. The network makes one error after each rule switch (black vertical lines) and quickly recovers its performance.
d. Performance as a function of trial number after a rule switch, for the same example network as in **c**.

Two rule attractor states in the PFC module maintained by interactions between modules

We first dissected the PFC module, which was trained to output the correct rule. Since there are two rules in the WCST task we used, we expected that the PFC module might have two attractor states corresponding to the two rules. Therefore, we examined the attractor structure in the dynamical landscape of the PFC module by initializing the network at states chosen randomly from the trial, and evolving the network autonomously (without any input) for 500 time steps (which equals 5 seconds in real time). Then, the dynamics of the PFC module during this evolution was visualized by applying principal component analysis to the population activity. The PFC population activity settles into one of two different attractor states depending on the rule that the initial state belongs to (Supplementary figure 2a). Therefore, there are two attractors in the dynamical landscape of the PFC module, corresponding to the two rules.

Historically, persistent neural activity corresponding to attractor states were first discovered in the PFC [27, 28, 29, 30]. However, more recent experiments found persistent neural activity in multiple brain regions, suggesting that long-range connections between brain regions may be essential for generating persistent activity [31, 32, 33, 34]. Inspired

by these findings, we wondered if the PFC module in our network could support the two rule attractor states by itself, or that the long-range connections between the PFC and the sensorimotor module are necessary to support them. To this end, we lesioned the inter-modular connections in the trained networks and repeated the simulation. Interestingly, we found that in the lesioned networks the PFC activity settled into a trivial fixed point corresponding to an inactive state (Supplementary Figure 2b). This result shows that the two rule attractor states in our model are dependent on the interactions between the PFC and the sensorimotor modules.

Two emergent subpopulations of excitatory neurons in the PFC module

For the PFC module to keep track of the current rule in effect, the module should stay in the same rule attractor state after receiving positive feedback, whereas transition to the other rule attractor state after receiving negative feedback. We reasoned that this network function might be mediated by single neurons that are modulated by the task rule and negative feedback, respectively. Therefore, we set out to look for these single neurons.

In the PFC module of the trained networks, there are indeed neurons whose activity is modulated by the task rule in a sustained fashion (example neurons in Supplementary Figure 1 and Figure 2a, top). In contrast, there are also neurons that show transient activity only after negative feedback. Furthermore, this activity is also rule-dependent. In other words, their activity is conjunctively modulated by negative feedback and the task rule (example neurons in Supplementary Figure 1, red traces and Figure 2a, bottom). We termed these two classes of neurons "rule neurons" and "conjunctive error x rule neurons" respectively.

We identified all the rule neurons and conjunctive error x rule neurons in the PFC module using a single neuron selectivity measure (see Methods for details). The two classes of neurons are clearly separable on the two-dimensional plane in Figure 2c, where the x axis is the input weight for negative feedback, and the y axis is the rule modulation, which is simply the difference in the mean activity between the two rules (for trials following a correct trial). As shown in Figure 2c, rule neurons receive negligible input about negative feedback, and many of them have activity modulated by rule. On the other hand, conjunctive error x rule neurons receive a substantial amount of input about negative feedback, yet their activity is minimally modulated by rule on trials following a correct trial (Figure 2b). This pattern was preserved when aggregating across trained networks (Figure 2c and Supplementary Figure 3). Interestingly, neurons with similar tuning profiles have been reported in the PFC and posterior parietal cortex of macaque monkeys performing the same WCST analog [8, 9]. Therefore, two classes of excitatory neurons emerged in the PFC module as a result of training.

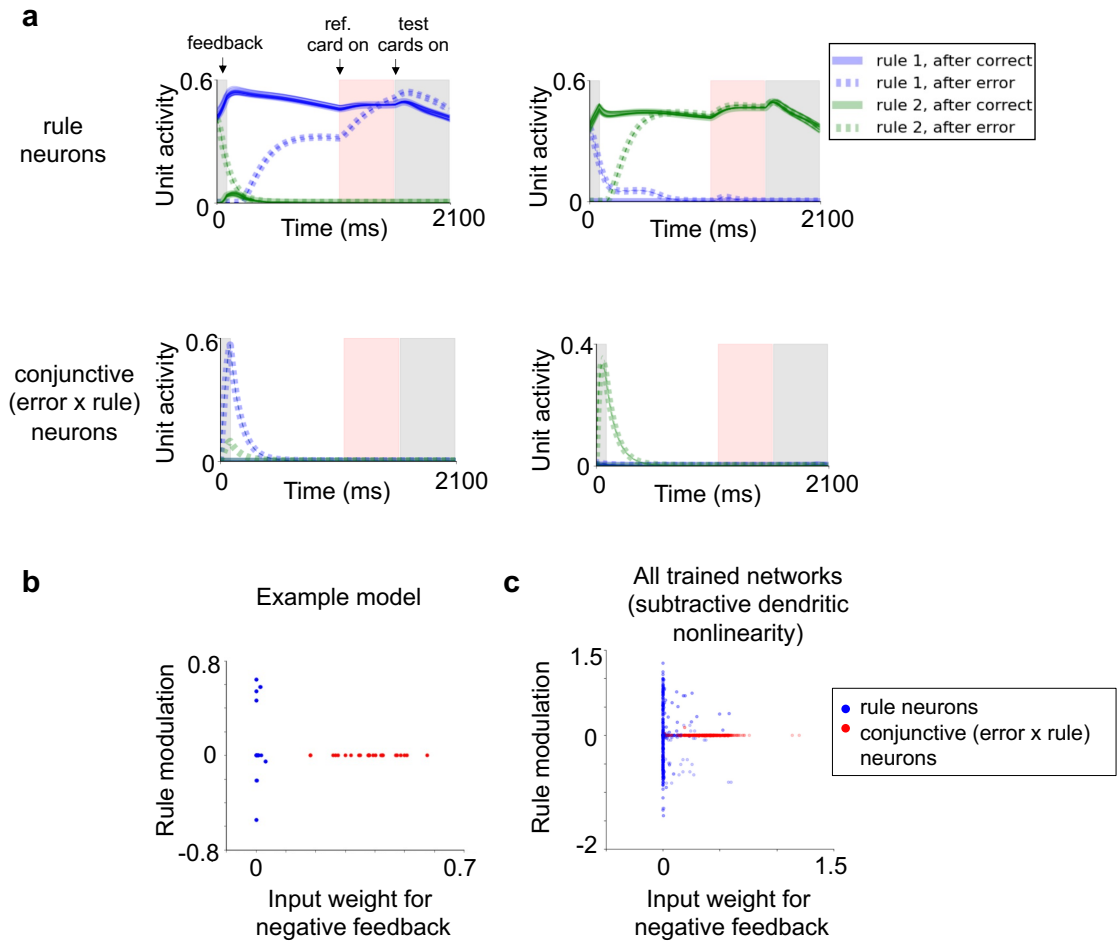


Figure 2: **Emergence of two subpopulations of excitatory neurons in the PFC module after training.**

a. Two example rule neurons (top) and conjunctive error x rule neurons (bottom). The thick solid traces represent the mean activity after correct trials, when those trials belong to color rule (blue) or shape rule (green) blocks. The thick dashed traces represent the mean activity after error trials, when those trials belong to color rule (blue) or shape rule (green) blocks. The thin traces represent single-trial activity. We use rule 1 and color rule, as well as rule 2 and shape rule interchangeably hereafter.

b. Rule neurons and conjunctive neurons are separable. The x axis represents the input weight for negative feedback, and the y axis is the difference between the mean activity over color rule trials and shape rule trials (for trials following a correct trial). As shown, the rule neurons (blue points) receive little input about negative feedback, but their activity is modulated by rule; The conjunctive error x rule neurons (red points) receive substantial input about negative feedback, but their activity is not modulated by rule (during trials following a correct trial).

c. The trend in **b** is preserved across a collection of trained networks. Here the results are shown for networks with subtractive dendritic nonlinearity. Networks with divisive dendritic nonlinearity show similar result (Supplementary Figure 3).

Maintaining and switching rule states via structured connectivity patterns between subpopulations of neurons within the PFC module

Given the existence of rule neurons and conjunctive error x rule neurons, what is the connectivity between them that enables the PFC module to stay in the same rule attractor state when receiving correct feedback, and switch to the other rule attractor state when receiving negative feedback?

To this end, we examined the connectivity between different subpopulations of excitatory neurons and PV neurons in the PFC module explicitly, by computing the mean connection strength between each pair of subpopulations. This analysis reveals that the excitatory rule neurons and PV rule neurons form a classic winner-take-all network architecture [35] with selective inhibitory populations [36, 37], where excitatory neurons preferring the same rule are more strongly connected, and also more strongly project to PV neurons preferring the same rule (Figure 3a). On the other hand, PV neurons project more strongly to both excitatory neurons and other PV neurons with the opposite rule preference (Figure 3a). This winner-take-all network motif together with the excitatory drive from the sensorimotor module (Supplementary Figure 2) is able to sustain one of the two attractor states.

Next, how are the rule neurons connected with the conjunctive error x rule neurons such that the sub-network formed by rule neurons can switch from one attractor to the other in the presence of the negative feedback input? Using the same method, we found that the connectivity between the rule neurons and the conjunctive error x rule neurons exhibit an interesting structure: the excitatory rule neurons more strongly target conjunctive error x rule neurons that prefer the opposite rule; the PV rule neurons more strongly target conjunctive error x rule neurons that prefer the same rule (Figure 3b, top two panels). On the other hand, the conjunctive error x rule neurons more strongly target the excitatory and PV rule neurons that prefer the same rule (Figure 3b, bottom two panels).

This connectivity structure gives rise to a simple circuit diagram of the PFC module (Figure 3c), which leads to an intuitive picture of the circuit mechanism underlying the switching of rule attractor state. For example, suppose the network is in the attractor state corresponding to color rule, and has just received a negative feedback and is about to switch to the attractor corresponding to the shape rule (Figure 3e, left). As shown in Figure 2b-c, the input current that represents the negative feedback mainly targets the conjunctive error x rule neurons. In addition, since the network is in the color rule state, the excitatory and PV neurons that prefer the color rule are more active than those that prefer the shape rule. According to Figure 3b (top two panels), the excitatory neurons that prefer the color rule strongly excite the error x shape rule neurons, and the PV neurons that prefer the color rule strongly inhibit the error x color rule neurons. Therefore, the error x shape rule neurons receive stronger total input than the error x color rule neurons, and will be more

active (Figure 3e, middle). Their activation will in turn excite the excitatory neurons and PV neurons that prefer the shape rule (Figure 3b, bottom two panels). Finally, due to the winner-take-all connectivity between the rule populations (Figure 3a), the excitatory and PV neurons that prefer the color rule will be suppressed, and the network will transition to the attractor state for the shape rule (Figure 3e, right).

Is the simplified circuit diagram (Figure 3c) consistent across trained networks, or different trained networks use different solutions? To examine this question, we computed a "connectivity bias" measure between each pair of populations for each trained network. This measure is greater than zero if the connectivity structure between a pair of neuron populations is closer to the one in the simplified circuit diagram in Figure 3c than to the opposite (see Methods for details). Across trained networks, we found that the connectivity biases were mostly greater than zero (Figure 3d), indicating that the same circuit motif for rule maintenance and switching underlies the PFC module across different trained networks.

Top-down propagation of the rule information through structured long-range connections

Given that the PFC module can successfully maintain and update the rule representation, how does it use the rule representation to reconfigure the sensorimotor mapping? The PFC module exerts top-down control through the monosynaptic pathway from the excitatory neurons in the PFC module to the excitatory neurons in the sensorimotor module, the tri-synaptic pathway that goes through the VIP and SST neurons in the sensorimotor module, as well as the di-synaptic pathway mediated by the PV neurons in the sensorimotor module (Figure 1b). We found that there are structured connectivity patterns along all three pathways. Along the monosynaptic pathway, excitatory neurons in the PFC module preferentially send long-range projections to the excitatory neurons in the sensorimotor module that prefer the same rule (Figure 4a). Along the tri-synaptic pathway, PFC excitatory neurons also send long-range projections to the SST and VIP interneurons in the sensorimotor module that prefer the same rule (Figure 4b-c). The SST neurons in turn send stronger inhibitory connections to the dendrite of the excitatory neurons in the sensorimotor module that prefer the opposite rule (Figure 4d). The PV neurons are also targeted by PFC excitatory neurons that prefer the same rule (Figure 4e), and they inhibit local excitatory neurons that prefer the opposite rule (Figure 4f). These trends are preserved across trained networks (Supplementary Figure 5a-f). Therefore, rule information is communicated to the sensorimotor module synergistically via the mono-synaptic excitatory pathway, the tri-synaptic pathway that involves the SST and VIP neurons, as well as the di-synaptic pathway that involves the PV neurons.

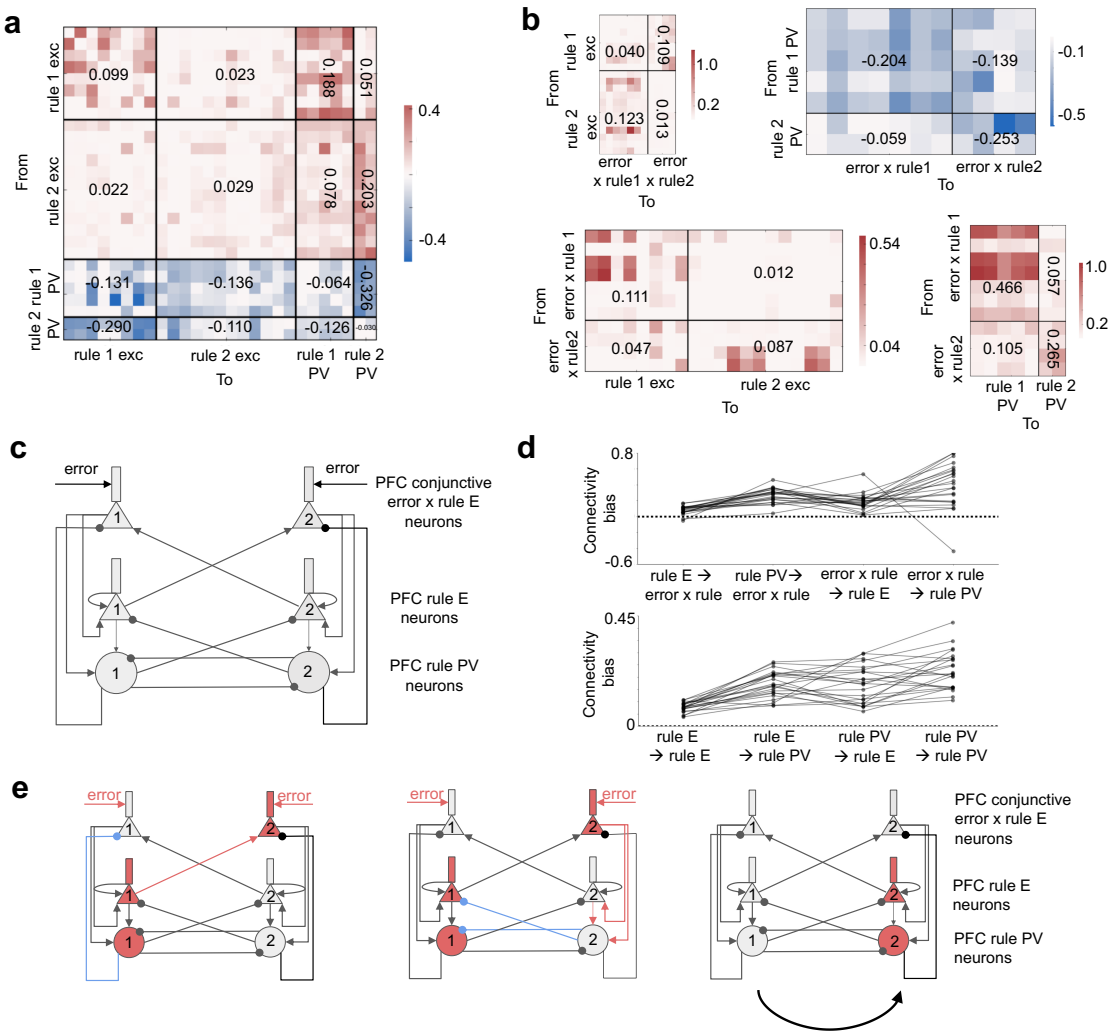


Figure 3: (Caption next page.)

A geometric property of the input weights enables rule-dependent action selection in the sensorimotor module

How does the sensorimotor module implement the sensorimotor transformation (from the cards to the spatial choice) given the top-down rule information from the PFC module? We sought to identify the structures in the input, recurrent and output connections of the sensorimotor module that give rise to this function. We started by classifying the excitatory neurons in the sensorimotor module according to the response locations that they prefer (Methods). Expectedly, we found that neurons preferring a given response location send stronger output connections to the output node that corresponds to the same response location (Figure 5a).

Next, we examined if there are structures in the pattern of the input weights that cor-

Figure 3: (Previous page.) **An emergent circuit wiring diagram in the PFC module enables un-cued switching between rule attractor states.**
a. The connectivity matrix between different populations of rule neurons, for an example model. Text indicates the mean connection strength between two populations. The excitatory rule neurons project more strongly to, and receive more input from, neurons with the same preferred rule. The PV rule neurons project more strongly to and receive more input from neurons with the opposite rule preference. As a result, rule neurons form a classic winner-take-all connectivity with selective inhibitory populations that maintain the two rule attractor state.
b. The connectivity between rule neurons and conjunctive error x rule neurons, for an example model. Top left: excitatory rule neurons project more strongly to the conjunctive error x rule neurons that prefer the opposite rule; Top right: PV rule neurons project more strongly to conjunctive error x rule neurons that prefer the same rule; Bottom left: conjunctive error x rule neurons project more strongly to the excitatory rule neurons that prefer the same rule; Bottom right: conjunctive error x rule neurons project more strongly to the PV rule neurons that prefer the same rule.
c. The simplified circuit diagram between rule neurons and conjunctive neurons based on the result of **b**. The weaker connections are ignored. Here, rule 1 represents the color rule and rule 2 represents the shape rule
d. A connectivity bias was computed to describe the extent to which the connectivity pattern between each pair of subpopulations conform to the simplified diagram in **c**. A value greater than 0 indicates the connectivity structure is more similar to that in **c** than to the opposite. The connectivity biases across all trained models are mostly above 0, both for the connection among rule neurons (top) and the connection between rule neurons and conjunctive error x rule neurons (bottom). Here the results are shown for networks with subtractive dendritic nonlinearity. Networks with divisive dendritic nonlinearity show similar result (Supplementary Figure 4).
e. A schematic showing how the simplified circuit can switch from the rule 1 attractor state to the rule 2 attractor state after receiving the input about negative feedback. The conjunctive error x rule 2 neurons receive excitation from the currently-active rule 1 excitatory neurons (red arrow, left panel), and the conjunctive error x rule 1 neurons receive inhibition from the currently-active rule 1 PV neurons (blue arrow, left panel). This makes conjunctive error x rule 2 neurons more active than conjunctive error x rule 1 neurons, even the negative feedback input is unbiased (left panel). The conjunctive error x rule 2 neurons then excite the rule 2 excitatory and PV neurons (red arrows, middle), which suppress the rule 1 excitatory and PV neurons due to the winner-take-all connectivity (blue arrows, middle) and eventually become more active (right). See text for details.

relates with the response selectivity of individual neurons. According to the rule of WCST, neurons selective for a particular rule and response location should receive the greatest input when the card at that location matches the reference card for the rule feature (color or shape). However, we found that neurons do not simply receive more external input from the input nodes that represent the card at their preferred response locations (Figure 5c),

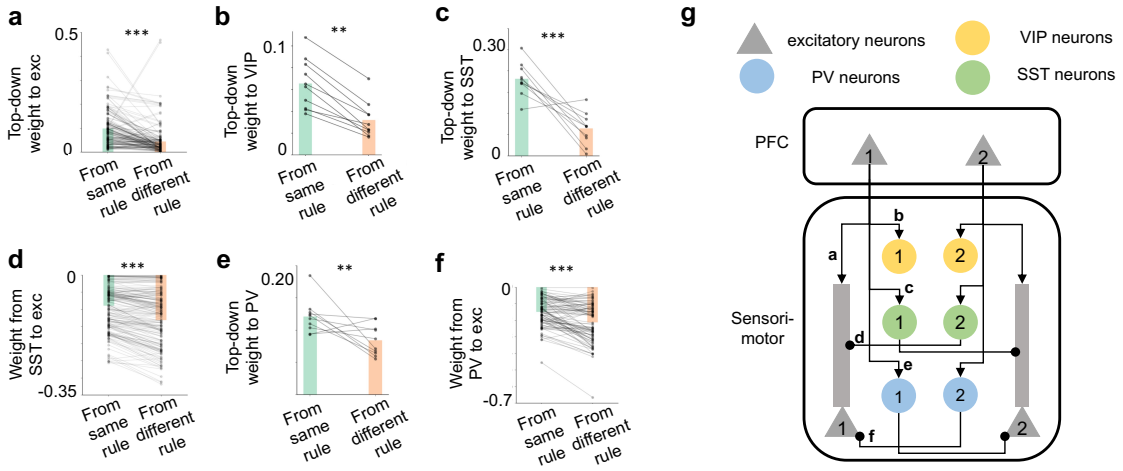


Figure 4: **Structured top-down connections enable the propagation of the rule information.**

a. Each line represents the mean connection strength onto one excitatory neuron in the sensorimotor module, from the PFC excitatory neurons that prefer the same rule and the different rule. Bars represent mean across neurons. PFC excitatory neurons project more strongly to sensorimotor excitatory neurons that prefer the same rule (Student's t test, $p < .001$).

b. Each line represents the mean connection strength onto one VIP neuron in the sensorimotor module, from the PFC excitatory neurons that prefer the same rule and the different rule. Bars represent mean across neurons. PFC excitatory neurons project more strongly to sensorimotor VIP neurons that prefer the same rule (Student's t test, $p = .002$).

c. Each line represents the mean connection strength onto one SST neuron in the sensorimotor module, from the PFC excitatory neurons that prefer the same rule and the different rule. Bars represent mean across neurons. PFC excitatory neurons project more strongly to sensorimotor SST neurons that prefer the same rule (Student's t test, $p < .001$).

d. Each line represents the mean connection strength onto one excitatory neuron of the sensorimotor module, from the sensorimotor SST neurons that prefer the same rule and the different rule. Bars represent mean across neurons. Sensorimotor SST neurons project more strongly to sensorimotor excitatory neurons that prefer the opposite rule (Student's t test, $p < .001$).

e. Each line represents the mean connection strength onto one PV neuron in the sensorimotor module, from the PFC excitatory neurons that prefer the same rule and the different rule. Bars represent mean across neurons. PFC excitatory neurons project more strongly to sensorimotor PV neurons that prefer the same rule (Student's t test, $p = 0.004$).

f. Each line represents the mean connection strength onto one excitatory neuron of the sensorimotor module, from the sensorimotor PV neurons that prefer the same rule and the different rule. Bars represent mean across neurons. Sensorimotor PV neurons project more strongly to sensorimotor excitatory neurons that prefer the opposite rule (Student's t test, $p < .001$).

g. The structure of the top-down connections as indicated by the results in **a-f**. The weaker connections are not shown.

Results in **a-f** are shown for networks with subtractive dendritic nonlinearity. Networks with divisive dendritic nonlinearity show similar result (Supplementary Figure 5).

indicating that a more intricate structure may be present in the input connections.

Let us first consider a simpler task without the rule component, so that analytical calculation can be used to gain some insights. In this simplified task, subjects simply need to choose the test card that matches *exactly* the reference card. Therefore, response location-selective neurons should receive the greatest amount of input when the card at their preferred response location matches *exactly* the reference card. Denote the input vectors representing the test cards and the reference card as $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, $\mathbf{x}_{\text{ref}}$, and their associated input weight vectors to a particular dendritic branch as $\mathbf{w}_1$, $\mathbf{w}_2$, $\mathbf{w}_3$, $\mathbf{w}_{\text{ref}}$ (Figure 5b). The total input onto this dendritic branch is therefore

$$I = \mathbf{w}_{\text{ref}} \cdot \mathbf{x}_{\text{ref}} + \mathbf{w}_1 \cdot \mathbf{x}_1 + \mathbf{w}_2 \cdot \mathbf{x}_2 + \mathbf{w}_3 \cdot \mathbf{x}_3. \tag{1}$$

Without loss of generality, let us assume that the preferred response location of this neuron is location 1. As mentioned previously, we would like the total input in Equation 1 to be large when the card at location 1 matches exactly the reference card, i.e. $\mathbf{x}_1 = \mathbf{x}_{\text{ref}}$. Equation 1 then simplifies to

$$I = (\mathbf{w}_{\text{ref}} + \mathbf{w}_1) \cdot \mathbf{x}_{\text{ref}} + \mathbf{w}_2 \cdot \mathbf{x}_2 + \mathbf{w}_3 \cdot \mathbf{x}_3. \tag{2}$$

For I to be large, the input weight vector from the input nodes representing the reference card ($\mathbf{w}_{\text{ref}}$) and the ones that represent the card at location 1 ($\mathbf{w}_1$) should align with each other (Figure 5d). This is due to the the basic mathematical property that the magnitude of the sum of two vectors is large when the two vectors align with each other (Figure 5d).

Inspired by this mathematical analysis, we asked if the input weight vectors of the trained networks obey this property. To this end, we computed the cosine similarity between $\mathbf{w}_{\text{ref}}$ and $\mathbf{w}_1$, $\mathbf{w}_2$, $\mathbf{w}_3$ in the trained networks, across all excitatory neurons in the sensorimotor module. Figure 5d shows the joint distribution of the three cosine similarities. Neurons were colored according to their preferred response locations. As can be seen in Figure 5e, although the values of cosine similarity are heterogeneous across neurons, they tend to be the largest for the input weight vector corresponding to the neuron's preferred response location (Figure 5e). This ensures that neurons receive maximal input when the card at their preferred response locations matches the reference card.

The above structure of the input weights allows neurons to detect the match between two cards, and in turn enables insights into the action selection mechanism during the full WCST task. In the WCST, a neuron needs to receive the greatest input when the card at its preferred response location matches the reference card *in the neuron's preferred rule feature* (color or shape). A simple way of achieving this is to have larger weights from the input nodes that represent its preferred rule (Figure 5f). In order to examine whether this

is the case in the trained networks, we computed the difference between the weights from the input nodes representing the color of the cards and those representing the shape of the cards, across all excitatory neurons in the sensorimotor module. These difference values are heterogenous among dendritic branches (Figure 5g, top), but they positively correlate with the neurons' rule selectivity (Figure 5g, bottom). Since a positive rule selectivity represents preference to the color rule (Methods), this indicates that excitatory neurons in the sensorimotor module receive greater inputs from input nodes representing their preferred rule features.

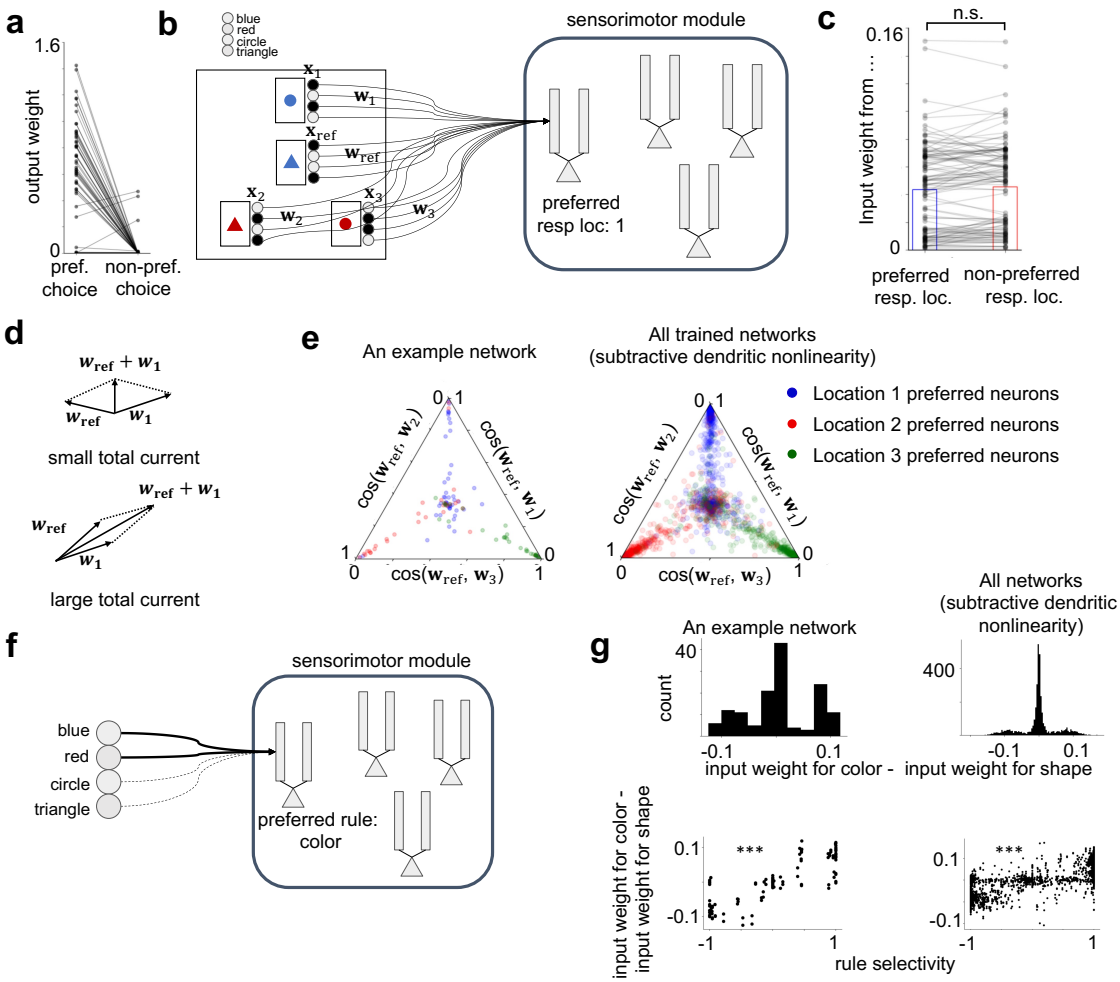


Figure 5: (Caption next page.)

Recurrent connectivity and dynamics within the sensorimotor module

Given that different populations of neurons in the sensorimotor module receive differential inputs about the external sensory stimuli via the structured input weights, how are they

Figure 5: (Previous page.) **Structures in the input and output weights of the sensorimotor module enable rule-dependent action selection.**
a. The weights to the output node that represents the preferred response location of each neuron are stronger than those to the output nodes that represent the non-preferred choices.
b. We examined the structure in the weights from the external input nodes to the excitatory neurons in the sensorimotor module. The card at each location on the screen is represented by a four-dimensional binary vector, where the two halves are one-hot vectors representing the color and shape of the card ([1010] for blue circle, [1001] for blue triangle, [0110] for red circle, [0101] for red triangle). Denote the input pattern for the reference card and three test cards on the screen as $\mathbf{x}_{\text{ref}}$, $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$ and the input weights associated with them (to a given dendritic branch) as $\mathbf{w}_{\text{ref}}$, $\mathbf{w}_1$, $\mathbf{w}_2$, $\mathbf{w}_3$.
c. Across the excitatory neurons in the sensorimotor module, the strength of the input connections from the sensory nodes that represent the card at each neuron's preferred response location is not significantly different from the input weights from the sensory nodes that represent the cards at other locations (Wilcoxon signed-rank test, $p = 0.99$).
d. Without loss of generality, assume location 1 is a particular neuron's preferred response location. It follows from the mathematical analysis (see text) that for that neuron, $\mathbf{w}_{\text{ref}}$ should be more aligned with $\mathbf{w}_1$ than with $\mathbf{w}_2$ or $\mathbf{w}_3$.
e. Left: The joint distribution of the cosine similarity between $\mathbf{w}_{\text{ref}}$ and $\mathbf{w}_1$, $\mathbf{w}_2$, $\mathbf{w}_3$. Blue, red, green points represent neurons that prefer response locations 1, 2 and 3, respectively. Note how neurons that prefer location i tend to have higher cosine similarity between $\mathbf{w}_{\text{ref}}$ and $\mathbf{w}_i$. Right: The joint distribution across all trained models. Data shown for networks with subtractive dendritic nonlinearity. Networks with divisive dendritic nonlinearity show similar result (Supplementary Figure 6).
f. Appropriate rule-dependent action selection requires that neurons receive stronger input connections from sensory nodes representing features of the cards that match the neurons' preferred rule (color or shape). Here is an illustration of the hypothetical input weight pattern onto a neuron that prefers the color rule. Solid arrows repreesnt stronger connections.
g. Top: the distribution of the difference between the input weights from the sensory nodes that represent the color and shape features, across the dendritic branches of all excitatory neurons in the sensorimotor module. Bottom: the difference value against the rule selectivity at the soma, across dendritic branches of all excitatory neurons in the sensorimotor module. The positive correlation shows that neurons receive more input from the input nodes that represent their preferred rule feature. Result is shown for one example model (left, Pearson's $R = 0.81$, $p < .001$) and across all models with subtractive dendritic nonlinearity (right, Pearson's $R = 0.70$, $p < .001$). Networks with divisive dendritic nonlinearity show similar result (Supplementary Figure 6).

recurrently connected to produce dynamics that lead to a categorical choice? To this end, we first visualized the population neural dynamics in the sensorimotor module by projecting them onto principal components (Figure 6a-b). As shown in Figure 6a, neural trajectories

during the inter-trial interval is clustered according to the task rule. During the response period, the neural trajectories are separable according to the response locations, albeit only in higher-order principal components (Figure 6b). In addition, the subspaces spanned by neural trajectories of different rules and response locations are more orthogonal to each other compared to randomly shuffled data (Figure 6c-d, Methods).

What connectivity structure gives rise to this signature in the population dynamics? To this end, we examined the pattern of connection strength between excitatory and PV neurons that prefer different rules and response locations. Figure 6e-f show example connectivity matrices after sorting the entries according to cell type and preferred rule or response location. As shown, the connectivity between different rule-selective or response location-selective populations partially conform to the winner-take-all structure, where excitatory neurons receive more input from and project more strongly to other neurons with the same preferred rule or response location, and the inhibitory neurons receive more input from and project more strongly to other neurons with different preferred rule or response locations (Figure 6e-f). Across networks, we computed connectivity biases between different rule and response location-selective populations analogously to that in the PFC module (Figure 3d, Methods). Many of the connectivity biases were significantly above zero (Figure 6g, h). This indicates that the excitatory and PV neurons in the sensorimotor module overall form a winner-take-all connectivity. In sum, the structured input (Figure 5e, g) and recurrent (Figure 6g, h) connections enable the sensorimotor module to implement the correct sensorimotor transformation given the top-down rule information.

SST neurons are essential to dendritic top-down gating

It has been observed that different dendritic branches of the same neuron can be tuned to different task variables [38, 39, 40, 41]. This property may enable individual dendritic branches to control the flow of information into the local network [17, 20]. Given these previous findings, we examined the coding of the top-down rule information at the level of individual dendritic branches. Since each excitatory neuron in our networks are modeled with two dendritic compartments, we examined the encoding of rule information by different dendritic branches of the same excitatory neuron in the sensorimotor module.

One strategy gating is for different dendritic branches of the same neuron to prefer the same rule, in which case these neurons form distinct populations that are recruited under different task rules (population-level gating, Figure 7a, right). An alternative strategy is for different dendritic branches of the same neuron to prefer different rules, which would enable these neurons to be involved in both task rules (dendritic branch-specific gating, Figure 7a, left).

In light of this, we examine for our trained networks to what extent different dendritic

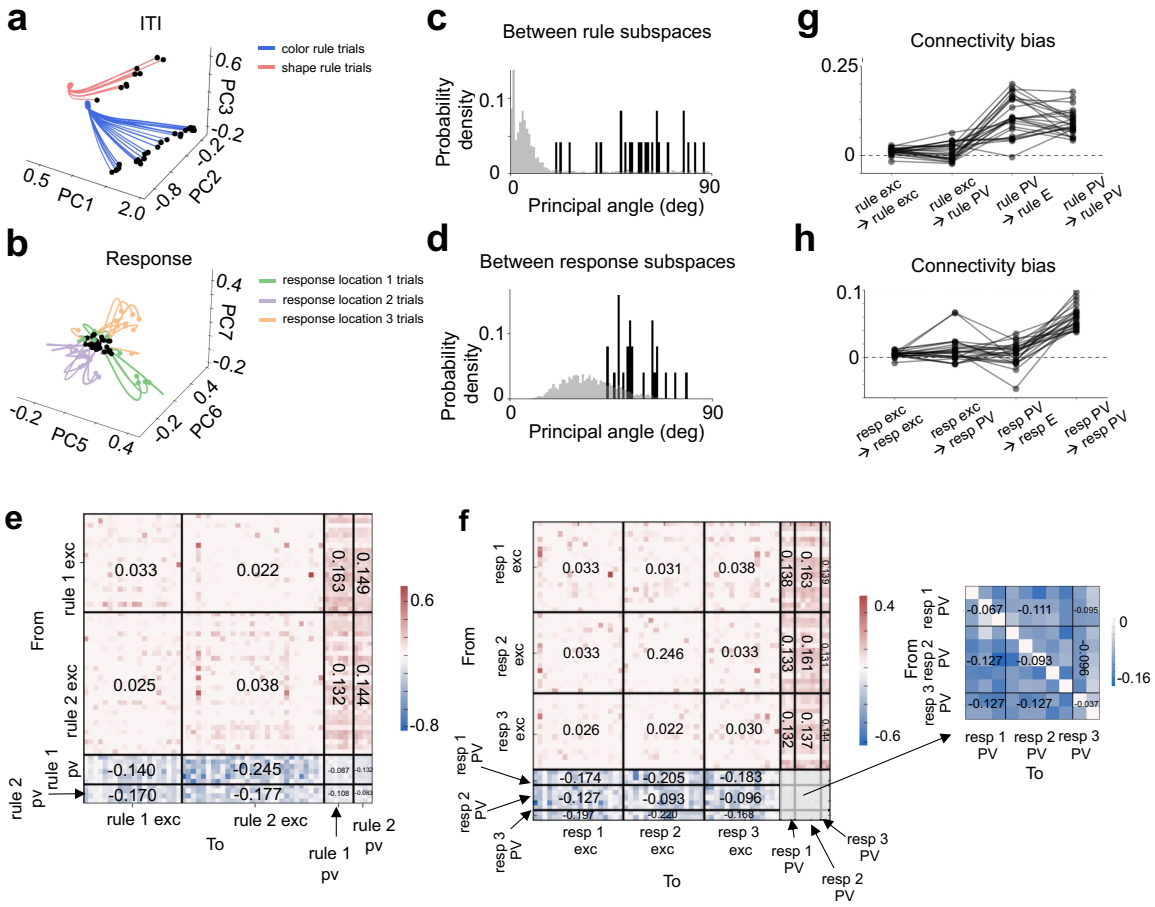


Figure 6: **Recurrent dynamics and connectivity within the sensorimotor module.**
a. Neural trajectories during the intertrial interval for different task rules, visualized in the space spanned by the first three principal components. Black circles represent the end of trajectories.
b. Neural trajectories during the response period for different choices, visualized in the space spanned by higher order principal components. Black circles represent the end of trajectories.
c. The principal angle between the subspaces spanned by neural trajectories during different task rules (gray distribution represents the principal angle obtained through shuffled data, see Methods). Each data point represents one trained network.
d. The principal angle between the subspaces spanned by neural trajectories during different choices (gray distribution represents the principal angle obtained through shuffled data, see Methods). Each data point represents one trained network.
e. The recurrent weight matrix between the excitatory neurons and PV neurons in the sensorimotor module. Neurons are sorted according to their preferred rules.
f. The same as **e** but with neurons sorted according to their preferred response locations.
g. The connectivity biases between different rule-selective populations across models.
h. The same as **g** but for different response-selective populations.
Data in **c**, **d**, **g**, **h** are for networks with subtractive dendritic nonlinearity. Networks with divisive dendritic nonlinearity show similar result (Supplementary Figure 7).

branches of the same neuron are selective for different task rules. When we plotted the rule selectivity of the two dendritic branches that belong to the same neuron, we found that the rule selectivity are highly correlated (Figure 7b). This indicates that the trained networks are mostly consistent with the population-level gating scenario, where different dendritic branches of the same neuron encode the same rule.

What factors might determine the extent to which the trained networks adopt these two solutions? Previous modeling work suggests that sparse connectivity from SST neurons to the dendrites of the excitatory neurons increases the degree of dendritic branch-specific gating, in the case where the connectivity is random [20]. To see if the same effect is present in our task-optimized network with structured connectivity, we re-trained networks with different levels of sparsity from 0 to 0.8 and studied its effect on the dendritic branch specificity of rule coding (Methods). We found that the degree of dendritic branch-specific encoding of the task rule increased with sparsity (see Figure 7c, d for subtractive dendritic nonlinearity; Supplementary Figure 8a for divisive dendritic nonlinearity). Therefore, the trained networks adopted a mixture of population-level and dendritic-level gating strategies for top-down control, and the balance between the two strategies depends on the sparsity of the connections from the SST neurons to the dendrites of excitatory neurons.

Indeed, SST neurons play an essential role in relaying the top-down rule information into the sensorimotor network and reconfiguring its dynamics according to the task rule. We simulated optogenetic inhibition by silencing the SST neurons in the sensorimotor module, which significantly impaired task performance (Figure 7e, see Methods section for details of the protocol). In addition, the principal angle between the subspaces for color rule and motion rule (Figure 6c) significantly decreased after SST neurons in the sensorimotor module were silenced (Figure 7f). Silencing of the SST neurons in the sensorimotor module also significantly diminished nonlinear mixed-selective coding of rule and stimulus among the excitatory neurons in the sensorimotor module (Figure 7g, Supplementary Figure 9, Methods), which has been proposed to be important for rule-based sensorimotor associations [42, 43, 44]. Taken together, these results highlight the role that SST neurons in the sensorimotor module play during top-down control. This analysis also shows that by combining artificial neural network with knowledge from neurobiology, it is possible to probe the functions of fine-scale biological components in cognitive behaviors.

Discussion

In this paper, we analyzed recurrent neural networks trained to perform a classic task involving un-cued task switching - the Wisconsin Card Sorting Test. The networks consist of a “PFC” module trained to represent the rule and interacts with a “sensorimotor” module that instantiates different sensorimotor mappings depending on the rule. In order to

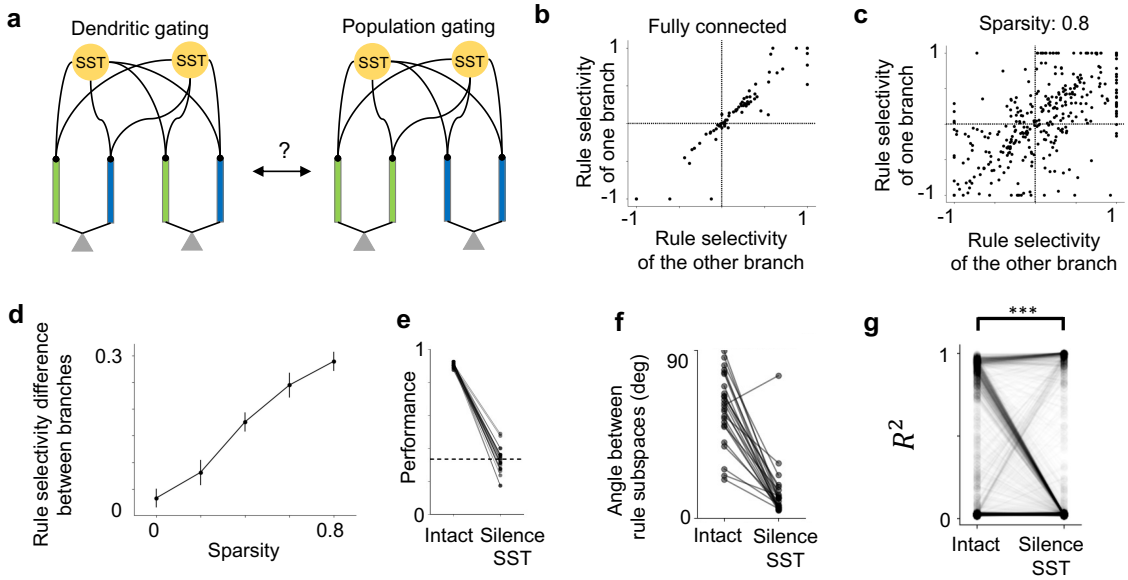


Figure 7: **Examining the role of SST neurons in top-down gating.**
a. Two scenarios for top-down gating. Blue and green color represent dendritic branches that prefer one of the two rules. Different dendritic branches of the same neuron could have similar (right) or different (left) rule selectivity.
b. The rule selectivity of one dendritic branch against the other, aggregated across all models where the connections from the SST neurons to the excitatory neurons are all-to-all. The rule selectivity for different dendritic branches of the same neuron are highly correlated.
c. The rule selectivity of one dendritic branch against the other, aggregated across all models where the 80% of the connections from the SST neurons to the excitatory neurons are frozen at 0 throughout training. Note the the rule selectivity for different dendritic branches of the same neuron are less correlated than in **b**.
d. The degree of dendritic branch-specific encoding of the task rule is quantified as the difference in the rule selectivity between the two dendritic branches of the same excitatory neuron in the sensorimotor module. Across all dendritic branches, this quantity increases with the sparsity of the SST $\rightarrow$ dendrite connectivity.
e. Task performance drops significantly after silencing SST neurons in the sensorimotor module. Each line represents a trained network.
f. The principal angle between rule subspaces (c.f. Figure 6c) drops significantly after silencing SST neurons in the sensorimotor module. Each line represents a trained network.
g. The strength of conjunctive coding of rule and stimulus (as measured by the R^2 value in a linear model with conjunctive terms, see Methods) decreases after silencing SST neurons in the sensorimotor module (Student's t-test, $p < .001$). Each line represents one neuron. Results are aggregated across networks.
Results in **b-g** are for networks with subtractive dendritic nonlinearity. Networks with divisive dendritic nonlinearity show similar result (Supplementary Figure 8).

study the functions of dendritic computation and different neuronal types, each module is endowed with excitatory neurons with two dendritic branches as well as three major types of inhibitory neurons - PV, SST and VIP. After training, we dissected the trained networks to elucidate a number of intra-areal and inter-areal neural circuit mechanisms underlying WCST, as summarized in Figure 8.

Mapping between model components and brain regions

Different components of the trained network can be mapped to different brain regions (Figure 8). The rule neurons and conjunctive error x rule neurons correspond to the putative functions of different subregions of PFC. While single neurons in the dorsal-lateral PFC are shown to encode the task rule [45], neurons in the anterior cingulate cortex are thought to be important for performance monitoring [46], and have been shown to receive more input about the feedback [47, 48, 49, 50]. The input to the PFC module about negative feedback may come from subcortical areas such as the amygdala [51] or from the dopamine neurons in the substantia nigra pars compacta (SNc) and ventral tegmental area (VTA) [52, 53]. The sensorimotor module may correspond to parietal cortex or basal ganglia which have been shown to be involved in sensorimotor transformations [54, 55]. The neurons in the input layer that encode the color and shape of the card stimuli exist in higher visual areas such as the inferotemporal cortex [56, 57, 58]. The neurons in the output layer that encode different response locations could correspond to movement location-specific neurons in the motor cortex [59].

Attractor states supported by inter-areal connections.

We observed two rule attractor states in the dynamical landscape of the networks supported by the interaction between the two modules (Supplementary Figure 2). This is contrary to the traditional notion that solely local interactions within the frontal regions are responsible for the maintenance of the information in working memory [27, 28, 29, 30]. Instead it suggests the possibility that the interactions between distributed brain regions underlie temporally-extended cognitive functions [31, 33, 60].

Circuit mechanism in the frontal-parietal network for rule maintenance and update

Two distinct types of responses among the excitatory neurons emerge in the PFC module as a result of training: neurons that only encode the rule, and neurons that conjunctively encode negative feedback and rule. Neurons that show conjunctive selectivity for rule and negative feedback have been reported in monkey prefrontal and parietal cortices while

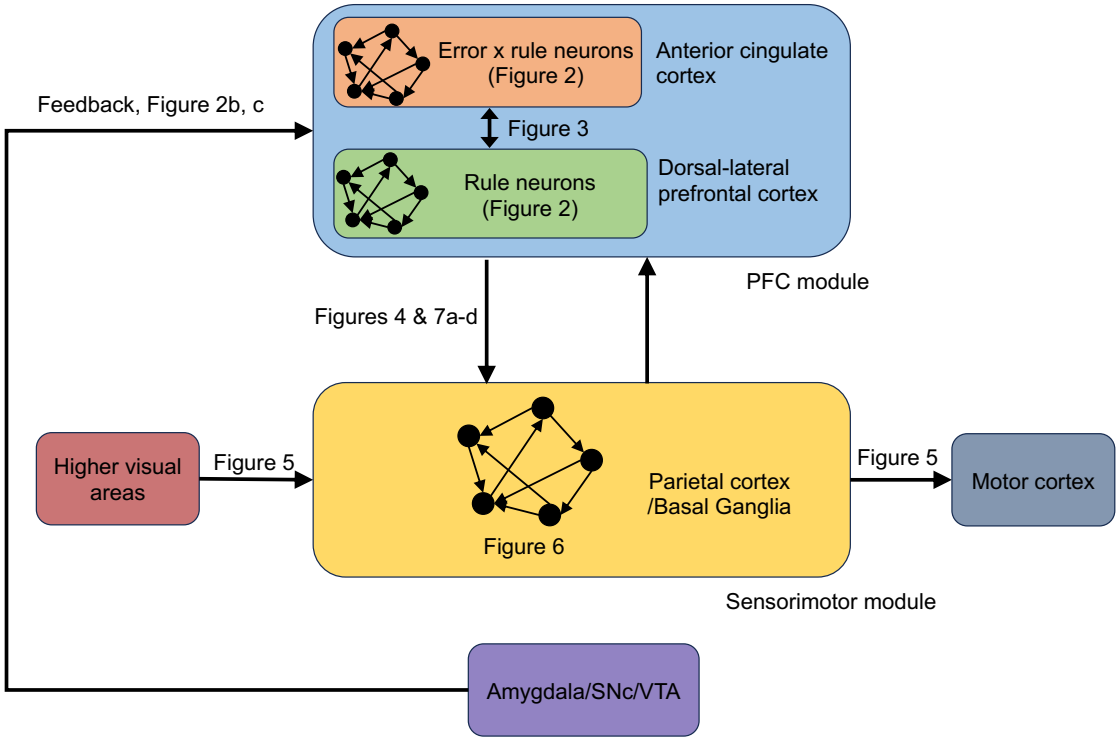


Figure 8: **A summary of the main results.**
Different components of the model can be mapped to different brain regions; The conjunctive error x rule neurons may reside in the anterior cingulate cortex; The rule neurons may be found in the dorsal-lateral PFC; The input to the PFC module about negative feedback may come from subcortical areas such as the amygdala or the midbrain dopamine neurons; The sensorimotor module may correspond to parietal cortex or basal ganglia which have been shown to be involved in sensorimotor transformations; Neurons in the input layer that encode the color and shape of the card stimuli exist in higher visual areas such as the inferotemporal cortex; Neurons in the output layer that encode different response locations could correspond to neurons in the motor cortex. See text for evidence that support these correspondences.

they perform the same WCST task [8, 9] . Theoretical work suggests that these mixed-selective neurons are essential for un-cued task switching because the network needs to switch between *different* rule attractor states after receiving the *same* input that signals negative feedback [61].

We further revealed the connectivity pattern between different populations of excitatory and PV neurons in the PFC module in order for the network to switch between rule attractor states (Figure 3c). In addition, this connectivity pattern is consistent across dozens of trained networks with different initializations and dendritic nonlinearities (Figure 3d and Supplementary Figure 4). This circuit mechanism bears resemblance to a previous circuit model of WCST [62]. In that model, the switching between different rule states is achieved by synaptic desensitization caused by the convergence of two signals - one that signals the recent activation of the synapse, and another that signals the negative feedback. As such, it does not predict the existence of neurons with conjunctive coding of negative feedback and rule, which has been observed experimentally [8, 9].

The simplified circuit for the PFC module in Figure 3c can be applied not only to rule switching, but to the switching between other behavioral states as well. For example, it resembles the head-direction circuit in fruit fly [63], where the offset in the connections between the neurons coding for head direction and those coding for the conjunction of angular velocity and head direction enables the update of the head-direction attractor state by the angular velocity input. In addition, this circuit structure may underlie the transition from staying to switching during patch foraging behavior. Indeed, in a laboratory task mimicking natural foraging for monkeys, it was found that neurons in the anterior cingulate cortex increase their firing rates to a threshold before animals switch to another food resource [64].

Connecting subspace to circuits

Methods that describe the representation and dynamics on the neuronal population level have gained increasing popularity and generated novel insights that cannot be discovered using single neuron analysis (e.g. [59, 65]). In the meantime, it would be valuable to connect population-level phenomena to their underlying circuit basis [66]. In our model, we found that silencing of the SST neurons has a specific effect on the population-level representation, namely, it decreases the angle between rule subspaces (Figure 7f). We also found that silencing the other types of inhibitory neurons has different effects (data not shown). Silencing the PV neurons leads to an instability of the network dynamics, whereas silencing the VIP neurons causes an insignificant decrease of the network performance. The lack of effect after silencing the VIP neurons is due to the fact that the VIP neurons are inhibited by the SST neurons in our model. Future work could study the function of VIP neurons under different connectivity constraints between SST and VIP neurons.

Evidence accumulation across trials during rule switching

Since our networks were optimized for task performance, they only made an error for the first trial after each rule switch. In other words, all errors were due to the incorrect rule representation. In reality, subjects make errors not only because they do not have the correct rule representation, but also because of errors in executing the sensorimotor mapping despite potentially correct rule representation. Consequently, negative feedback should not necessarily induce the rule switching behavior. It has been reported that monkeys performing un-cued rule switching accumulate negative feedback over multiple trials as the evidence that a rule switch has occurred [67, 10, 68]. In addition, the amount of evidence accumulated for each negative feedback depends on the subject's confidence for that trial. The higher the confidence of an error trial is, the more likely that the animal switches its rule representation in the subsequent trials, consistent with a Bayesian ideal observer model [67, 10, 68]. From this view, our networks effectively implement high confidence for every trial, and therefore switch to the other rule whenever they receive negative feedback. Future work can train networks to reproduce the subjects' behavior as closely as possible rather than to maximize performance, in order to study how negative feedback is combined with other internal signals across multiple trials to infer the relevant task rule.

WCST with more than two rules

The WCST task that our networks are trained on involves only two rules, whereas the WCST deployed clinically usually has three rules. In this case, it cannot be guaranteed that the network switches to the correct new rule after making only one error. A straightforward extension of our simplified PFC module (Figure 3c) to more than two rules would result in the network going back to the wrong rule attractor state that has just been visited more than one trial back during switching (for example, when required to switch from rule 1 to rule 3, the network might try rule 1, rule 2, and rule 1 again). This is because the activation of the conjunctive error x rule neurons is transient (c.f. Figure 2a bottom). Therefore the network does not have memory about errors that occurred more than one trial before. We speculate that this behavior can be mitigated by incorporating neuronal processes with time constants on the order of several trials, such as the activation of the metabotropic glutamate receptor [69, 70, 71]. Future work could incorporate biological processes with longer time constants into the network model in order to investigate their functions in WCST.

In conclusion, our approach of incorporating neurobiological knowledge into training RNNs can provide a fruitful way to build circuit models that are functional, high-dimensional, and reflect the heterogeneity of biological neural networks. Dissecting these networks can make useful cross-level predictions that connect biological ingredients with circuit mechanisms and cognitive functions.

Acknowledgements: This work was supported by James Simons Foundation Grant 543057SPI, the National Institutes of Health grant R01MH062349, and the ONR grant N00014-23-1-2040. YL thanks Aldo Battista, Vishwa Goudar, Lucas Tian, Kenneth Kay, Haohong Li's lab, Yu Qi's lab, Sage Chen, Yi Sun's lab, Jianguang Ni's lab, Bo Shen, Shuo Chen, Xiaohan Zhang and all members of Xiao-Jing Wang's lab for helpful discussions and comments on the manuscript.

Data availability statement: All computer code used to generate the model and results will be uploaded to Github upon acceptance of the manuscript.

Methods

Model setup

The RNN consists of two bidirectionally-connected modules, the PFC module and the sensorimotor module. Each module consists of 70 excitatory neurons and 30 inhibitory neurons. Each excitatory neuron has 2 dendritic compartments. The inhibitory neurons are evenly divided into three types: PV, SST and VIP. Different types of neurons have different connectivity, inspired by experimental findings [72]: PV neurons target the somatic compartment of excitatory neurons and other PV neurons, SST neurons target the dendritic compartment of excitatory neurons as well as PV and VIP neurons, and VIP neurons target SST neurons. Excitatory neurons target other excitatory neurons, PV and SST neurons. The connection strength between all other types of neurons are fixed at zero throughout training.

Only excitatory neurons send long-range projections to other modules. The long-range projections from the sensorimotor module to the PFC module target the dendritic compartment of the excitatory neurons and the PV neurons. This is inspired by the experimental evidence that PV neurons mediate feedforward inhibition [12]. The long-range top-down projections from the PFC to the sensorimotor module target the dendritic compartments of the excitatory neurons and all three types of inhibitory neurons. Finally, external inputs to both modules target the dendritic compartment of excitatory neurons and PV neurons.

The dynamics of the somata of the excitatory neurons in the RNN are described by

$$\tau \frac{dh_{\text{soma, exc}}}{dt} = -h_{\text{soma, exc}} + f_{\text{soma, exc}}(W_{\text{rec,eff}} h_{\text{soma, exc}} + \sum_{\text{dendrites}} h_{\text{dendrite}}), \tag{3}$$

where $\tau = 100$ ms, $dt = 10$ ms. $f_{\text{soma, exc}}$ is the rectified linear activation function:

$$f_{\text{soma, exc}} = \begin{cases} 1 & x > 0 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

The effective connectivity matrix $W_{\text{rec, eff}}$ is given by

$$W_{\text{rec, eff}} = W_{\text{rec}} * M, \tag{5}$$

where M is the mask matrix consisting of 1, 0 and -1 according to the cell-type-specific connectivity described above. $*$ denotes element-wise product. Note that the dendritic compartments are coupled to their corresponding somatic compartments with a fixed coupling strength of 1.

The dendritic activity is a nonlinear function of the excitatory and inhibitory inputs.

$$h_{\text{dendrite}} = f_{\text{dendrite}}(I_e, I_i). \tag{6}$$

The functional form of f_{dendrite} is described in the next section.

The inhibitory neurons are modeled as standard point neurons

$$\tau \frac{dh_{\text{soma, inh}}}{dt} = -h_{\text{soma, inh}} + f_{\text{soma, inh}}(W_{\text{rec,eff}} h_{\text{soma, inh}} + W_{\text{in}} I_{\text{in}}). \tag{7}$$

The activation function $f_{\text{soma, inh}}$ is the same as the one used for excitatory neurons $f_{\text{soma, exc}}$ (Equation 4).

The readouts are only from the somata of excitatory neurons. We use simple linear readout.

$$y = W_{\text{out}} h_{\text{exc, soma}}. \tag{8}$$

There is one readout from each module. The readout from the PFC module is trained to match the correct rule, whereas the one from the sensorimotor module is trained to match the correct response location.

Variations in the model hyperparameters

Dendritic nonlinearities. We trained models with two types of dendritic nonlinearities f_{dendrite} - subtractive and divisive. They are inspired by in-vitro and computational studies showing different types of inhibitory modulation on the dendritic activity depending on the location of inhibition relative to excitation [23]. Both types of dendritic nonlinearities are sigmoidal functions of the excitatory input. Under subtractive nonlinearity, as the inhibitory input increases, the turning point of the sigmoid function moves to larger values, consistent with the experimental observation when the inhibitory current is injected at the same location or more distal than the excitation [23]. For the divisive nonlinearity, the turning point of the sigmoid is not affected by the level of inhibition, but the saturating level of the sigmoid function decreases with the level of inhibition, consistent with the experimental observation when the inhibitory current is injected close to the soma [23].

The equations of the different dendritic nonlinearities are given by:

$$f_{\text{subtractive}}(I_e, I_i) = \tanh\left(I_e - I_i\right)$$

$$f_{\text{divisive}}(I_e, I_i) = k_1(1 + \tanh(I_e - 1)) + k_2,$$

where $k_1 = \frac{1}{e^{I_i}}$ and $k_2 = -1 - \tanh(-1)$. The constant k_2 is introduced such that the value of the function is 0 when both excitatory and inhibitory inputs are 0.

Initializations. We initialize all the connectivity matrices either using a normal distribution with mean 0 and standard deviation $\sqrt{\frac{2}{N}}$ (where N is the total number of recurrent units) or a uniform distribution between $-\sqrt{\frac{6}{N}}$ and $\sqrt{\frac{6}{N}}$.

Sparsity of the SST→dendrite connectivity in the sensorimotor module. To study how the degree of dendritic branch-specific rule encoding in the sensorimotor module is affected by the sparsity of the connections from SST neurons to the dendrite of excitatory neurons, we varied this sparsity by fixing a fraction of randomly chosen weights to be 0 throughout training.

Random seeds. For each combination of the hyperparameter configuration introduced above (except the sparsity), we trained models using 50 random seeds for Pytorch (other random seeds were fixed). For each sparsity level other than 0, we trained models using 10 random seeds for Pytorch.

Task

The network was trained on an analog of the Wisconsin Card Sorting Test (WCST) used for monkeys [21, 8, 9]. Each trial starts with the presentation of a "reference card" for 500 ms, after which three "test cards" appear around the reference card for 500 ms. Each card contains an object with a specific color (out of two possible colors) and shape (out of two possible shapes). Among the three test cards, one of them matches the color of the reference card, another one matches the shape of the reference card, and the third card matches neither feature of the reference card. Depending on the rule (color or shape), the response location where the test card has the same color or shape feature as the reference card should be chosen. The choice should be made during the 500 ms when both the reference card and the test cards are presented. At the end of this period, a feedback signal is presented for 100 ms, indicating whether the choice is correct or incorrect. This is followed by a 1 second inter-trial interval.

The task rule switches after a random number of trials, without informing the network. Therefore, the network inevitably makes an error for the first trial after the rule switch since

it has not yet received the information that the rule has switched. The network should then adjust its behavior to the new rule by utilizing the feedback signal.

Representation of inputs and outputs

Each card is represented as a four-dimensional binary vector, where each entry represents the presence of one of the two colors and shapes. The feedback input is a two-dimensional one-hot vector, where the two entries represent positive and negative feedback. The target output for the sensorimotor module is a three-dimensional one-hot vector, where each entry represents one response location on the screen. This target is non-zero only during the 500 ms response period when both the reference card and the test cards are presented. The target output for the PFC module is a two-dimensional one-hot vector, where each entry represents one rule. This target is non-zero during the entire trial.

Training method

During training, the networks ran continuously across 20 consecutive trials with 3 random rule switches. The network dynamics were not reset during the inter-trial interval. The loss function was aggregated across the 20 trials.

$$L = \sum_{\text{trial}=1}^{20} \sum_{\text{t}} \left[(y_{\text{pfc, trial}}(t) - \hat{y}_{\text{rule, trial}}(t))^2 + (y_{\text{sensorimotor, trial}}(t) - \hat{y}_{\text{choice, trial}}(t))^2 \right] \quad (9)$$

The standard backpropogation through time algorithm [73] with the Adam optimizer [74] was used to update all the connection weights.

We also used curriculum learning to speed up training. Initially, the stimulus and choice of the previous trial together with the feedback were provided to the PFC module. After the network reached 85% performance, the input about the previous stimulus was removed. When the network reached 85% performance again, the information about the previous choice information was removed. The networks were then trained until they reached 95% performance.

Single neuron selectivity metric

The selectivity index (SI) for rule is defined as

$$SI_{\text{rule}} = \frac{h(\text{color}) - h(\text{shape})}{|h(\text{color})| + |h(\text{shape})|}, \quad (10)$$

where $h(\text{color})$ and $h(\text{shape})$ represent the trial-averaged single neuron activity during color rule and shape rule, respectively. Neural activity was first averaged over the inter-trial interval before further averaging across trials.

The error selectivity is defined similarly

$$SI_{\text{error}} = \frac{h(\text{after error}) - h(\text{after correct})}{|h(\text{after error})| + |h(\text{after correct})|}, \tag{11}$$

where $h(\text{after error})$ and $h(\text{after correct})$ are the mean single neuron activity after error and correct trials, respectively. Neural activity was first averaged across the feedback presentation and inter-trial interval periods before being averaged across trials.

The selectivity for response location is defined slightly differently. Since there are 3 responses, we computed one selectivity for each response. The response selectivity for choosing the test card at location X ($X = 1, 2, 3$) is defined as

$$SI_{\text{response}}(X) = \frac{h(X) - h(\bar{X})}{|h(X)| + |h(\bar{X})|}, \tag{12}$$

where $h(\bar{X})$ represents the mean single neuron activity across trials where the choice is not location X. Neural activity was averaged over the response period before being averaged across trials.

Classification criteria for different neuronal populations

Each neuron in the PFC module was classified as a "rule neuron" if the absolute value of its rule selectivity was greater than 0.5 and the absolute value of its error selectivity was smaller than 0.5. It was classified as an "error neuron" if the absolute value of its rule selectivity was smaller than 0.5 and its error selectivity was greater than 0.5. Error neurons with greater mean activity during the color rule trials that follow error trials were defined as error x color rule neurons, and the other error neurons were defined as error x shape rule neurons.

Each neuron in the sensorimotor module was classified as preferring one of the three response locations according to the response location when the activity of that neuron was the highest. Each unit was also classified as preferring one of the two rules according to the rule during which it had the higher average activity.

Connectivity bias

The connectivity bias (CB) was defined as the average weight difference between different sub-population of neurons. A positive value indicates an agreement with the simplified circuit diagram of the PFC module (Figures 3c) or a winner-take-all connectivity diagram for different response location-selective populations in the sensorimotor module. For example,

the connectivity bias from the PFC PV neurons to the PFC excitatory neurons is given by

$$\begin{aligned} CB(\text{PFC PV} \longrightarrow \text{PFC E}) = &\ \bar{W}(\text{PFC PV rule1} \longrightarrow \text{PFC E rule2}) \\ &+ \bar{W}(\text{PFC PV rule2} \longrightarrow \text{PFC E rule1}) \\ &- \bar{W}(\text{PFC PV rule1} \longrightarrow \text{PFC E rule1}) \\ &- \bar{W}(\text{PFC PV rule2} \longrightarrow \text{PFC E rule2}), \end{aligned} \tag{13}$$

where for example $\bar{W}(\text{PFC PV rule1} \longrightarrow \text{PFC E rule2})$ represents the average (unsigned) connection strength from the PFC PV neurons that prefer rule 1 to PFC excitatory neurons that prefer rule 2. Here rule 1 refers to color rule and rule 2 refers to shape rule.

The other connectivity biases were defined analogously.

$$\begin{aligned} CB(\text{PFC E} \longrightarrow \text{PFC E}) = &\ \bar{W}(\text{PFC E rule1} \longrightarrow \text{PFC E rule1}) \\ &+ \bar{W}(\text{PFC E rule2} \longrightarrow \text{PFC E rule2}) \\ &- \bar{W}(\text{PFC E rule1} \longrightarrow \text{PFC E rule2}) \\ &- \bar{W}(\text{PFC E rule2} \longrightarrow \text{PFC E rule1}) \end{aligned} \tag{14}$$

$$\begin{aligned} CB(\text{PFC E} \longrightarrow \text{PFC PV}) = &\ \bar{W}(\text{PFC E rule1} \longrightarrow \text{PFC PV rule1}) \\ &+ \bar{W}(\text{PFC E rule2} \longrightarrow \text{PFC PV rule2}) \\ &- \bar{W}(\text{PFC E rule1} \longrightarrow \text{PFC PV rule2}) \\ &- \bar{W}(\text{PFC E rule2} \longrightarrow \text{PFC PV rule1}) \end{aligned} \tag{15}$$

$$\begin{aligned} CB(\text{PFC PV} \longrightarrow \text{PFC PV}) = &\ \bar{W}(\text{PFC PV rule1} \longrightarrow \text{PFC PV rule2}) \\ &+ \bar{W}(\text{PFC PV rule2} \longrightarrow \text{PFC PV rule1}) \\ &- \bar{W}(\text{PFC PV rule1} \longrightarrow \text{PFC PV rule1}) \\ &- \bar{W}(\text{PFC PV rule2} \longrightarrow \text{PFC PV rule2}) \end{aligned} \tag{16}$$

$$\begin{aligned} CB(\text{PFC E} \longrightarrow \text{PFC error x rule}) = &\ \bar{W}(\text{PFC E rule1} \longrightarrow \text{PFC error x rule2}) \\ &+ \bar{W}(\text{PFC E rule2} \longrightarrow \text{PFC error x rule1}) \\ &- \bar{W}(\text{PFC E rule1} \longrightarrow \text{PFC error x rule1}) \\ &- \bar{W}(\text{PFC E rule2} \longrightarrow \text{PFC error x rule2}) \end{aligned} \tag{17}$$

$$\begin{aligned} CB(\text{PFC PV} \longrightarrow \text{PFC error x rule}) = {} & \bar{W}(\text{PFC PV rule1} \longrightarrow \text{PFC error x rule1}) \\ & + \bar{W}(\text{PFC PV rule2} \longrightarrow \text{PFC error x rule2}) \\ & - \bar{W}(\text{PFC PV rule1} \longrightarrow \text{PFC error x rule2}) \\ & - \bar{W}(\text{PFC PV rule2} \longrightarrow \text{PFC error x rule1}) \end{aligned} \tag{18}$$

$$\begin{aligned} CB(\text{PFC error x rule} \longrightarrow \text{PFC E}) = {} & \bar{W}(\text{PFC error x rule1} \longrightarrow \text{PFC E rule1}) \\ & + \bar{W}(\text{PFC error x rule2} \longrightarrow \text{PFC E rule2}) \\ & - \bar{W}(\text{PFC error x rule1} \longrightarrow \text{PFC E rule2}) \\ & - \bar{W}(\text{PFC error x rule2} \longrightarrow \text{PFC E rule1}) \end{aligned} \tag{19}$$

$$\begin{aligned} CB(\text{PFC error x rule} \longrightarrow \text{PFC PV}) = {} & \bar{W}(\text{PFC error x rule1} \longrightarrow \text{PFC PV rule1}) \\ & + \bar{W}(\text{PFC error x rule2} \longrightarrow \text{PFC PV rule2}) \\ & - \bar{W}(\text{PFC error x rule1} \longrightarrow \text{PFC PV rule2}) \\ & - \bar{W}(\text{PFC error x rule2} \longrightarrow \text{PFC PV rule1}) \end{aligned} \tag{20}$$

The connectivity biases within the sensorimotor module are defined as (SM is short for sensorimotor)

$$\begin{aligned} CB(\text{SM response E} \longrightarrow \text{SM response E}) = {} & \bar{W}(\text{SM E response 1} \longrightarrow \text{SM E response 1}) \\ & + \bar{W}(\text{SM E response 2} \longrightarrow \text{SM E response 2}) \\ & + \bar{W}(\text{SM E response 3} \longrightarrow \text{SM E response 3}) \\ & - \bar{W}(\text{SM E response 1} \longrightarrow \text{SM E response 2 and 3}) \\ & - \bar{W}(\text{SM E response 2} \longrightarrow \text{SM E response 1 and 3}) \\ & - \bar{W}(\text{SM E response 3} \longrightarrow \text{SM E response 1 and 2}). \end{aligned} \tag{21}$$

In the last equation, for example, $\bar{W}(\text{SM response 1} \longrightarrow \text{SM response 2 and 3})$ represents the mean connection strength from excitatory neurons in the sensorimotor module that prefer response location 1 to those that prefer response locations 2 and 3.

The other connectivity biases were defined similarly

$$\begin{aligned} CB(\text{SM response E} \longrightarrow \text{SM response PV}) = \bar{W}(&\text{SM E response 1} \longrightarrow \text{SM PV response 1}) \\ + \bar{W}(&\text{SM E response 2} \longrightarrow \text{SM PV response 2}) \\ + \bar{W}(&\text{SM E response 3} \longrightarrow \text{SM PV response 3}) \\ - \bar{W}(&\text{SM E response 1} \longrightarrow \text{SM PV response 2 and 3}) \\ - \bar{W}(&\text{SM E response 2} \longrightarrow \text{SM PV response 1 and 3}) \\ - \bar{W}(&\text{SM E response 3} \longrightarrow \text{SM PV response 1 and 2}). \end{aligned} \tag{22}$$

$$\begin{aligned} CB(\text{SM response PV} \longrightarrow \text{SM response E}) = \bar{W}(&\text{SM PV response 1} \longrightarrow \text{SM E response 2 and 3}) \\ + \bar{W}(&\text{SM PV response 2} \longrightarrow \text{SM E response 1 and 3}) \\ + \bar{W}(&\text{SM PV response 3} \longrightarrow \text{SM E response 1 and 2}) \\ - \bar{W}(&\text{SM PV response 1} \longrightarrow \text{SM E response 1}) \\ - \bar{W}(&\text{SM E response 2} \longrightarrow \text{SM PV response 2}) \\ - \bar{W}(&\text{SM E response 3} \longrightarrow \text{SM PV response 3}). \end{aligned} \tag{23}$$

$$\begin{aligned} CB(\text{SM response PV} \longrightarrow \text{SM response PV}) = \bar{W}(&\text{SM PV response 1} \longrightarrow \text{SM PV response 2 and 3}) \\ + \bar{W}(&\text{SM PV response 2} \longrightarrow \text{SM PV response 1 and 3}) \\ + \bar{W}(&\text{SM PV response 3} \longrightarrow \text{SM PV response 1 and 2}) \\ - \bar{W}(&\text{SM PV response 1} \longrightarrow \text{SM PV response 1}) \\ - \bar{W}(&\text{SM PV response 2} \longrightarrow \text{SM PV response 2}) \\ - \bar{W}(&\text{SM PV response 3} \longrightarrow \text{SM PV response 3}). \end{aligned} \tag{24}$$

Simulation of the optogenetic inhibition

Optogenetic inhibition was simulated by clamping the activity of neurons at 0 throughout the entire trial and the inter-trial interval.

Principal angle between subspaces

The principal angle between two subspaces is a generalization of angle between lines and planes in Euclidean space to arbitrary dimensions [75]. It can be computed by iteratively finding unit length "principal vectors" in the two subspaces that have the greatest inner

product, subject to the condition that the principal vectors are orthogonal to all previous principal vectors [76].

In computing the principal angles between different rule-selective and response-selective subspaces, we first determined the dimensionality of the subspaces using the participation ratio [77]. Then the principal angles were computed using the "subspace_angles" function from the Python package Scipy.

To obtain a shuffled distribution, we first evenly split all trials belonging to a particular rule or choice into two halves. Then, we generated two subspaces from neural trajectories during the two group of trials. A principal angle between these two subspaces was then computed for each rule and choice, and then averaged across all rules or choices to obtain a principal angle from shuffled data. This process was repeated 100 times to generate a distribution of principal angles from shuffled data.

Assessing the strength of non-linear mixed selectivity

The extent to which neurons in the sensorimotor module encode the conjunction of stimulus and rule in a non-linear fashion was evaluated using the coefficient of determination of a linear regression model. To tease apart non-linear and linear mixed selectivity, we first fitted the mean activity of each neuron during response period using a set of regressors that represent either the rule or the stimulus alone:

$$FR(n, tr) = \sum_{s} \beta_{n,s} \mathbb{1}(\text{stim}(tr) = s) + \sum_{r} \beta_{n,r} \mathbb{1}(\text{rule}(tr) = r), \tag{25}$$

where $FR(n, tr)$ is the firing rate of neuron n during trial tr. $\mathbb{1}$ is the indicator function. For example, $\mathbb{1}(\text{stim}(tr) = s) = 1$ if the stimulus during trial tr is s, and it is 0 otherwise.

Then, another linear regression model was fitted on the residual activity unexplained by the linear regression model above, using the conjunction of rule and stimulus as regressors:

$$\tilde{FR}(n, tr) = \sum_{s,r} \beta_{n,s,r} \mathbb{1}(\text{stim}(tr) = s, \text{rule}(tr) = r), \tag{26}$$

where $\tilde{FR}(n, tr)$ is the firing rate of neuron n during trial tr minus the predicted firing rate from the model equation 25. The R^2 value of this regression model was used to represent the strength of non-linear mixed selectivity.

References

[1] David A Grant and Esta Berg. A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a weigl-type card-sorting problem. *Journal of experimental psychology*, 38(4):404, 1948.

[2] Brenda Milner. Effects of different brain lesions on card sorting: The role of the frontal lobes. *Archives of neurology*, 9(1):90–100, 1963.

[3] R Dias, TW Robbins, and AC Roberts. Primate analogue of the wisconsin card sorting test: effects of excitotoxic lesions of the prefrontal cortex in the marmoset. *Behavioral neuroscience*, 110:872, 1996.

[4] RE Passingham. Non-reversal shifts after selective prefrontal ablations in monkeys (macaca mulatta). *Neuropsychologia*, 10(1):41–46, 1972.

[5] Katsuyuki Sakai. Task set and prefrontal cortex. *Annu. Rev. Neurosci.*, 31:219–245, 2008.

[6] M. J. Buckley, F. A. Mansouri, H. Hoda, M. Mahboubi, P. G. Browning, S. C. Kwok, A. Phillips, and K. Tanaka. Dissociable components of rule-guided behavior depend on distinct medial and prefrontal regions. *Science*, 325:52–58, 2009.

[7] E. K. Miller and J. D. Cohen. An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, 24:167–202, 2001.

[8] Farshad A Mansouri, Kenji Matsumoto, and Keiji Tanaka. Prefrontal cell activities related to monkeys' success and failure in adapting to rule changes in a wisconsin card sorting test analog. *Journal of Neuroscience*, 26(10):2745–2756, 2006.

[9] Tsukasa Kamigaki, Tetsuya Fukushima, and Yasushi Miyashita. Cognitive set reconfiguration signaled by macaque posterior parietal neurons. *Neuron*, 61(6):941–951, 2009.

[10] Morteza Sarafyazd and Mehrdad Jazayeri. Hierarchical reasoning by neural circuits in the frontal cortex. *Science*, 364(6441):eaav8911, 2019.

[11] Takuya Ito, Guangyu Robert Yang, Patryk Laurent, Douglas H Schultz, and Michael W Cole. Constructing neural network models from brain data reveals representational transformations linked to adaptive behavior. *Nature communications*, 13(1):673, 2022.

[12] Kristen Delevich, Jason Tucciarone, Z Josh Huang, and Bo Li. The mediodorsal thalamus drives feedforward inhibition in the anterior cingulate cortex via parvalbumin interneurons. *Journal of Neuroscience*, 35(14):5743–5753, 2015.

[13] Hyun-Jae Pi, Balázs Hangya, Duda Kvitsiani, Joshua I Sanders, Z Josh Huang, and Adam Kepecs. Cortical interneurons that specialize in disinhibitory control. *Nature*, 503(7477):521–524, 2013.

[14] Siyu Zhang, Min Xu, Tsukasa Kamigaki, Johnny Phong Hoang Do, Wei-Cheng Chang, Sean Jenvay, Kazunari Miyamichi, Liqun Luo, and Yang Dan. Long-range and local circuits for top-down modulation of visual cortex processing. *Science*, 345(6197):660–665, 2014.

[15] William Muñoz, Robin Tremblay, Daniel Levenstein, and Bernardo Rudy. Layer-specific modulation of neocortical dendritic inhibition during active wakefulness. *Science*, 355(6328):954–959, 2017.

[16] Andreas J Keller, Mario Dipoppa, Morgane M Roth, Matthew S Caudill, Alessandro Ingrosso, Kenneth D Miller, and Massimo Scanziani. A disinhibitory circuit for contextual modulation in primary visual cortex. *Neuron*, 108(6):1181–1193, 2020.

[17] X. J. Wang, J. Tegnér, C. Constantinidis, and P. S. Goldman-Rakic. Division of labor among distinct subtypes of inhibitory neurons in a cortical microcircuit of working memory. *Proceedings of the National Academy of Science, USA*, 101(5):1368–73, 2004.

[18] A. Kepecs and G. Fishell. Interneuron cell types are fit to function. *Nature*, 505:318–326, 2014.

[19] R. Tremblay, S. Lee, and B. Rudy. GABAergic interneurons in the neocortex: From cellular properties to circuits. *Neuron*, 91:260–292, 2016.

[20] Guangyu Robert Yang, John D Murray, and Xiao-Jing Wang. A dendritic disinhibitory circuit mechanism for pathway-specific gating. *Nature communications*, 7(1):1–14, 2016.

[21] Kiyoshi Nakahara, Toshihiro Hayashi, Seiki Konishi, and Yasushi Miyashita. Functional mri of macaque monkeys performing a cognitive set-shifting task. *Science*, 295(5559):1532–1536, 2002.

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

[23] Monika Jadi, Alon Polsky, Jackie Schiller, and Bartlett W Mel. Location-dependent effects of inhibition on local spiking in pyramidal neuron dendrites. *PLoS computational biology*, 8(6):e1002550, 2012.

[24] Valerio Mante, David Sussillo, Krishna V Shenoy, and William T Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *nature*, 503(7474):78–84, 2013.

[25] Guangyu Robert Yang, Madhura R Joglekar, H Francis Song, William T Newsome, and Xiao-Jing Wang. Task representations in neural networks trained to perform many cognitive tasks. *Nature neuroscience*, 22(2):297–306, 2019.

[26] H Francis Song, Guangyu R Yang, and Xiao-Jing Wang. Training excitatory-inhibitory recurrent neural networks for cognitive tasks: a simple and flexible framework. *PLoS computational biology*, 12(2):e1004792, 2016.

[27] Joaquin M Fuster. Unit activity in prefrontal cortex during delayed-response performance: neuronal correlates of transient memory. *Journal of neurophysiology*, 36(1):61–78, 1973.

[28] Shintaro Funahashi, Charles J Bruce, and Patricia S Goldman-Rakic. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of neurophysiology*, 61(2):331–349, 1989.

[29] P. Goldman-Rakic. Cellular basis of working memory. *Neuron*, 14:477–85, 1995.

[30] Ranulfo Romo, Carlos D Brody, Adrián Hernández, and Luis Lemus. Neuronal correlates of parametric working memory in the prefrontal cortex. *Nature*, 399(6735):470–473, 1999.

[31] Thomas B Christophel, P Christiaan Klink, Bernhard Spitzer, Pieter R Roelfsema, and John-Dylan Haynes. The distributed nature of working memory. *Trends in cognitive sciences*, 21(2):111–124, 2017.

[32] Matthew L Leavitt, Diego Mendoza-Halliday, and Julio C Martinez-Trujillo. Sustained activity encoding working memories: not fully distributed. *Trends in Neurosciences*, 40(6):328–346, 2017.

[33] Zengcai V Guo, Hidehiko K Inagaki, Kayvon Daie, Shaul Druckmann, Charles R Gerfen, and Karel Svoboda. Maintenance of persistent activity in a frontal thalamocortical loop. *Nature*, 545(7653):181–186, 2017.

[34] Kartik K Sreenivasan and Mark D'Esposito. The what, where and how of delay activity. *Nature reviews neuroscience*, 20(8):466–481, 2019.

[35] Kong-Fatt Wong and Xiao-Jing Wang. A recurrent network mechanism of time integration in perceptual decisions. *Journal of Neuroscience*, 26(4):1314–1328, 2006.

[36] Farzaneh Najafi, Gamaleldin F Elsayed, Robin Cao, Eftychios Pnevmatikakis, Peter E Latham, John P Cunningham, and Anne K Churchland. Excitatory and inhibitory subnetworks are equally selective during decision-making and emerge simultaneously during learning. *Neuron*, 105(1):165–179, 2020.

[37] James P Roach, Anne K Churchland, and Tatiana A Engel. Choice selective inhibition drives stability and competition in decision circuits. *Nature Communications*, 14(1):147, 2023.

[38] Hongbo Jia, Nathalie L Rochefort, Xiaowei Chen, and Arthur Konnerth. Dendritic organization of sensory input to cortical neurons in vivo. *Nature*, 464(7293):1307–1312, 2010.

[39] Joseph Cichon and Wen-Biao Gan. Branch-specific dendritic ca2+ spikes cause persistent synaptic plasticity. *Nature*, 520(7546):180–185, 2015.

[40] Shannon K Rashid, Victor Pedrosa, Martial A Dufour, Jason J Moore, Spyridon Chavlis, Rodrigo G Delatorre, Panayiota Poirazi, Claudia Clopath, and Jayeeta Basu. The dendritic spatial code: branch-specific place tuning and its experience-dependent decoupling. *BioRxiv*, pages 2020–01, 2020.

[41] Jakob Voigts and Mark T Harnett. Somatic and dendritic encoding of spatial variables in retrosplenial cortex differs during 2d navigation. *Neuron*, 105(2):237–245, 2020.

[42] Mattia Rigotti, Omri Barak, Melissa R Warden, Xiao-Jing Wang, Nathaniel D Daw, Earl K Miller, and Stefano Fusi. The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497(7451):585–90, 2013.

[43] Atsushi Kikumoto and Ulrich Mayr. Conjunctive representations that integrate stimuli, responses, and rules are critical for action selection. *Proceedings of the National Academy of Sciences*, 117(19):10603–10608, 2020.

[44] Atsushi Kikumoto, Ulrich Mayr, and David Badre. The role of conjunctive representations in prioritizing and selecting planned actions. *Elife*, 11:e80153, 2022.

[45] Jonathan D Wallis, Kathleen C Anderson, and Earl K Miller. Single neurons in prefrontal cortex encode abstract rules. *Nature*, 411(6840):953–956, 2001.

[46] Matthew M Botvinick, Jonathan D Cohen, and Cameron S Carter. Conflict monitoring and anterior cingulate cortex: an update. *Trends in cognitive sciences*, 8(12):539–546, 2004.

[47] René Quilodran, Marie Rothe, and Emmanuel Procyk. Behavioral shifts and action valuation in the anterior cingulate cortex. *Neuron*, 57(2):314–325, 2008.

[48] Nils Kolling, Marco K Wittmann, Tim EJ Behrens, Erie D Boorman, Rogier B Mars, and Matthew FS Rushworth. Value, search, persistence and model updating in anterior cingulate cortex. *Nature neuroscience*, 19(10):1280–1285, 2016.

[49] Farshad Alizadeh Mansouri, David J Freedman, and Mark J Buckley. Emergence of abstract rules in the primate brain. *Nature Reviews Neuroscience*, 21(11):595–610, 2020.

[50] Timothy Spellman, Malka Svei, Jesse Kaminsky, Gabriela Manzano-Nieves, and Conor Liston. Prefrontal deep projection neurons enable cognitive flexibility via persistent feedback monitoring. *Cell*, 184(10):2750–2766, 2021.

[51] C Daniel Salzman and Stefano Fusi. Emotion, cognition, and mental state representation in amygdala and prefrontal cortex. *Annual review of neuroscience*, 33:173–202, 2010.

[52] Clay B Holroyd and Michael GH Coles. The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychological review*, 109(4):679, 2002.

[53] Masayuki Matsumoto and Okihide Hikosaka. Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature*, 459(7248):837–841, 2009.

[54] Richard A Andersen and He Cui. Intention, action planning, and decision making in parietal-frontal circuits. *Neuron*, 63(5):568–583, 2009.

[55] Bernard W Balleine and John P O'doherty. Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology*, 35(1):48–69, 2010.

[56] Bevil R Conway. Color vision, cones, and color-coding in the cortex. *The neuroscientist*, 15(3):274–290, 2009.

[57] Rosa Lafer-Sousa and Bevil R Conway. Parallel, multi-stage processing of colors, faces and shapes in macaque inferior temporal cortex. *Nature neuroscience*, 16(12):1870–1878, 2013.

[58] Le Chang, Pinglei Bao, and Doris Y Tsao. The representation of colored objects in macaque color patches. *Nature communications*, 8(1):2064, 2017.

[59] Mark M Churchland, John P Cunningham, Matthew T Kaufman, Justin D Foster, Paul Nuyujukian, Stephen I Ryu, and Krishna V Shenoy. Neural population dynamics during reaching. *Nature*, 487(7405):51–56, 2012.

[60] Nicholas A Steinmetz, Peter Zatka-Haas, Matteo Carandini, and Kenneth D Harris. Distributed coding of choice, action and engagement across the mouse brain. *Nature*, 576(7786):266–273, 2019.

[61] Mattia Rigotti, Daniel Ben Dayan Rubin, Xiao-Jing Wang, and Stefano Fusi. Internal representation of task rules by recurrent dynamics: the importance of the diversity of neural responses. *Frontiers in computational neuroscience*, 4:24, 2010.

[62] Stanislas Dehaene and Jean-Pierre Changeux. The wisconsin card sorting test: Theoretical analysis and modeling in a neuronal network. *Cerebral cortex*, 1(1):62–79, 1991.

[63] Daniel Turner-Evans, Stephanie Wegener, Herve Rouault, Romain Franconville, Tanya Wolff, Johannes D Seelig, Shaul Druckmann, and Vivek Jayaraman. Angular velocity integration in a fly heading circuit. *Elife*, 6:e23496, 2017.

[64] Benjamin Y Hayden, John M Pearson, and Michael L Platt. Neuronal basis of sequential foraging decisions in a patchy environment. *Nature neuroscience*, 14(7):933–939, 2011.

[65] João D Semedo, Amin Zandvakili, Christian K Machens, M Yu Byron, and Adam Kohn. Cortical areas interact through a communication subspace. *Neuron*, 102(1):249–259, 2019.

[66] Christopher Langdon, Mikhail Genkin, and Tatiana A Engel. A unifying perspective on neural manifolds and circuits for cognition. *Nature Reviews Neuroscience*, pages 1–15, 2023.

[67] Braden A Purcell and Roozbeh Kiani. Hierarchical decision processes that operate over distinct timescales underlie choice and changes in strategy. *Proceedings of the national academy of sciences*, 113(31):E4531–E4540, 2016.

[68] Cheng Xue, Lily E Kramer, and Marlene R Cohen. Dynamic task-belief is an integral part of decision-making. *Neuron*, 110(15):2503–2511, 2022.

[69] S Murray Sherman. The function of metabotropic glutamate receptors in thalamus and cortex. *The Neuroscientist*, 20(2):136–149, 2014.

[70] Michael E Hasselmo, Andrew S Alexander, Alec Hoyland, Jennifer C Robinson, Marianne J Bezaire, G William Chapman, Ausra Saudargiene, Lucas C Carstensen, and Holger Dannenberg. The unexplored territory of neural models: Potential guides for exploring the function of metabotropic neuromodulation. *Neuroscience*, 456:143–158, 2021.

[71] Chong Guo, Vincent Huson, Evan Z Macosko, and Wade G Regehr. Graded heterogeneity of metabotropic signaling underlies a continuum of cell-intrinsic temporal responses in unipolar brush cells. *Nature Communications*, 12(1):5491, 2021.

[72] Xiaolong Jiang, Shan Shen, Cathryn R Cadwell, Philipp Berens, Fabian Sinz, Alexander S Ecker, Saumil Patel, and Andreas S Tolias. Principles of connectivity among morphologically defined cell types in adult neocortex. *Science*, 350(6264):aac9462, 2015.

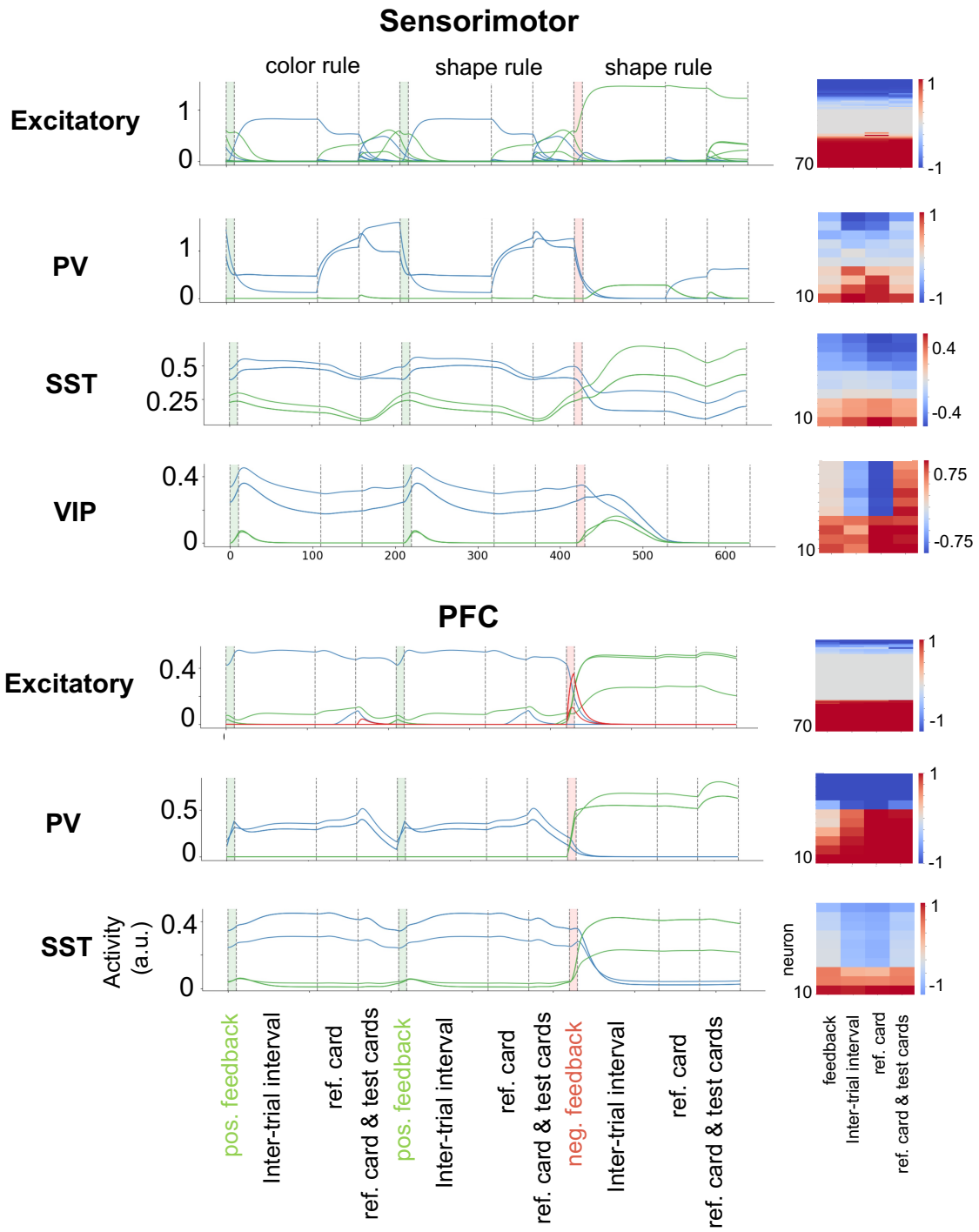
[73] Paul J Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990.

[74] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

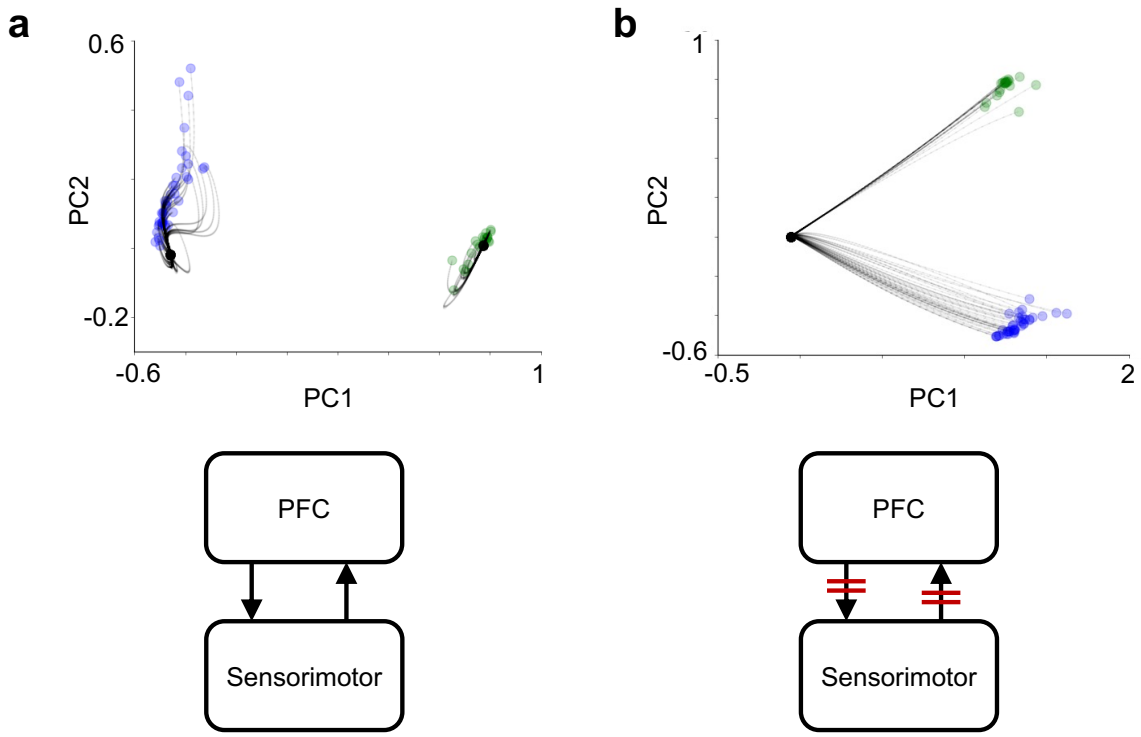
[75] Camille Jordan. Essai sur la géométrie à n dimensions. *Bulletin de la Société mathématique de France*, 3:103–174, 1875.

[76] ke Björck and Gene H Golub. Numerical methods for computing angles between linear subspaces. *Mathematics of computation*, 27(123):579–594, 1973.

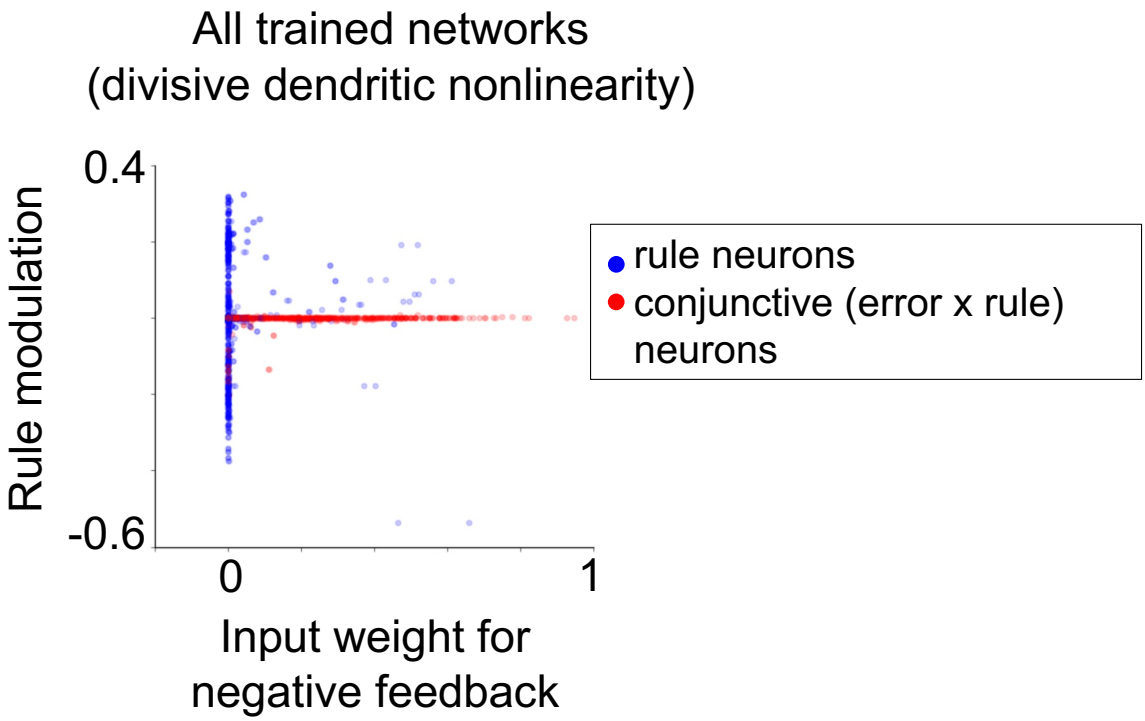
[77] Peiran Gao, Eric Trautmann, Byron Yu, Gopal Santhanam, Stephen Ryu, Krishna Shenoy, and Surya Ganguli. A theory of multineuronal dimensionality, dynamics and measurement. *BioRxiv*, page 214262, 2017.



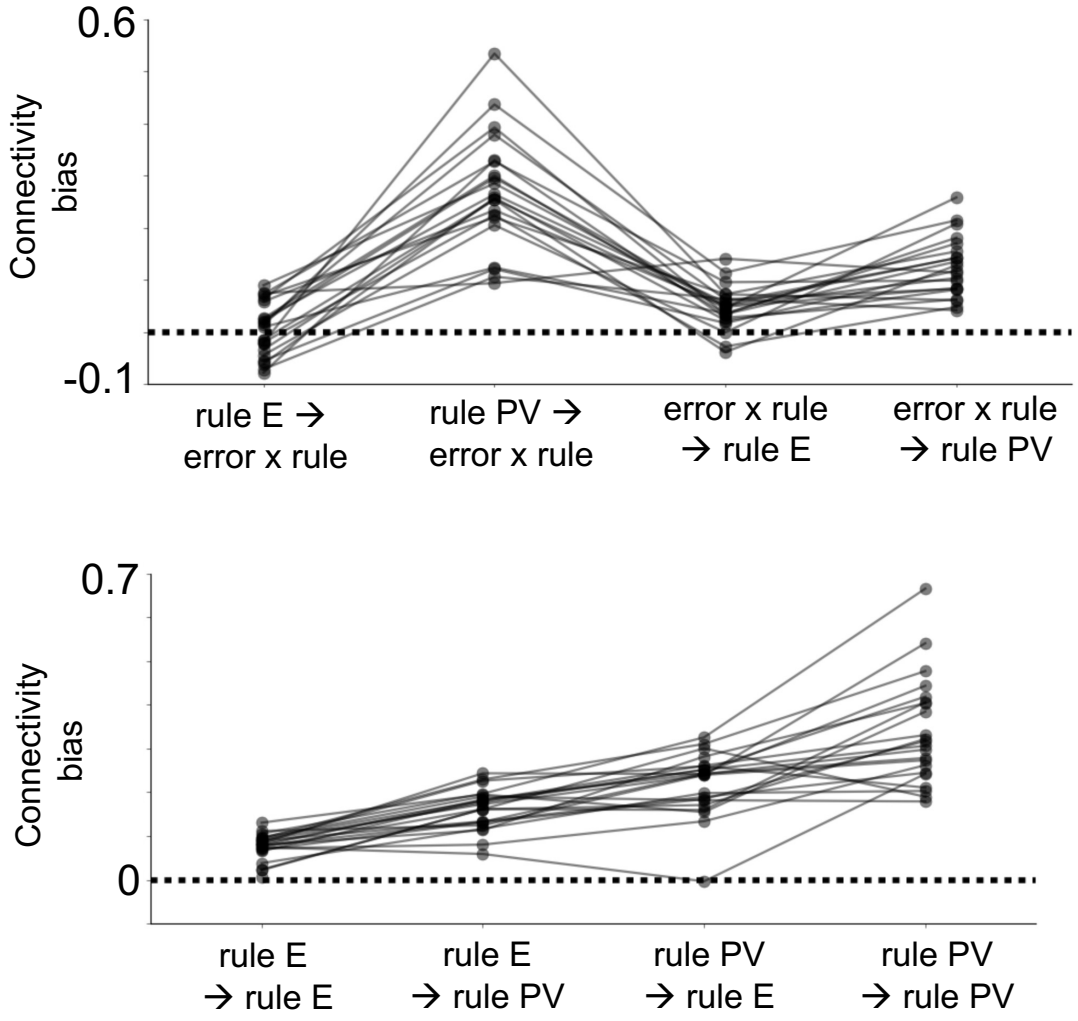
Supplementary Figure 1: **Activity of single neurons**.
Left: activity from example neurons across three consecutive trials with a rule switch. Blue and green traces represent neurons with higher activity during the color rule and shape rule respectively. Red traces for the PFC excitatory neurons represent neurons that are preferentially activated by negative feedback.
Right: Rule selectivity across neurons and task epochs.
VIP neurons in the PFC module do not receive any excitatory inputs, therefore are not shown.



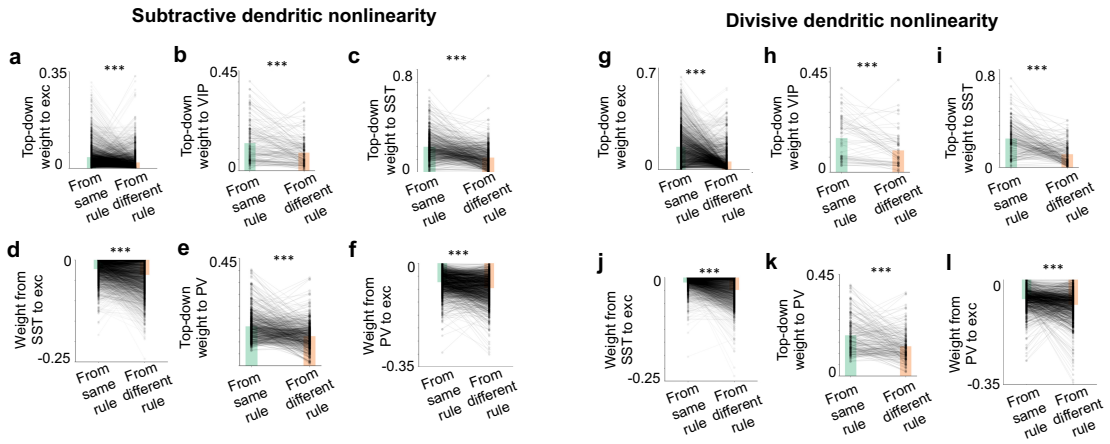
Supplementary Figure 2: **Two attractor states in the PFC module supported by inter-modular connections.**
The neural trajectories that represent the autonomous dynamics of the trained RNN with intact (**a**) or lesioned (**b**) inter-modular connections. The starting states are selected from a random time point during color rule (blue) and shape rule (green) trials. The black points represent the final state after 500 timesteps (5 seconds).



Supplementary Figure 3: **The emergence of two populations of excitatory neurons in the PFC module of networks with divisive dendritic nonlinearity.** The rule modulation against input weight for negative feedback for all the rule neurons and conjunctive error x rule neurons in the PFC module of networks with divisive dendritic nonlinearity (c.f. Figure 2c).



Supplementary Figure 4: The connectivity biases between functional subpopulations of the excitatory and PV neurons in the PFC module of networks with divisive dendritic nonlinearity (c.f. Figure 3d).



Supplementary Figure 5: **Structure in the top-down projections, across all networks.**

a. Each line represents the mean connection strength onto one excitatory neuron in the sensorimotor module, from the PFC excitatory neurons that prefer the same rule and the different rule. Bars represent the mean across neurons. PFC excitatory neurons project more strongly to sensorimotor excitatory neurons that prefer the same rule (c.f. Figure 4a).

b. Each line represents the mean connection strength onto one VIP neuron in the sensorimotor module, from the PFC excitatory neurons that prefer the same rule and the different rule. Bars represent the mean across neurons. PFC excitatory neurons project more strongly to sensorimotor VIP neurons that prefer the same rule (c.f. Figure 4b).

c. Each line represents the mean connection strength onto one PV neuron in the sensorimotor module, from the PFC excitatory neurons that prefer the same rule and the different rule. Bars represent the mean across neurons. PFC excitatory neurons project more strongly to sensorimotor PV neurons that prefer the same rule (c.f. Figure 4c).

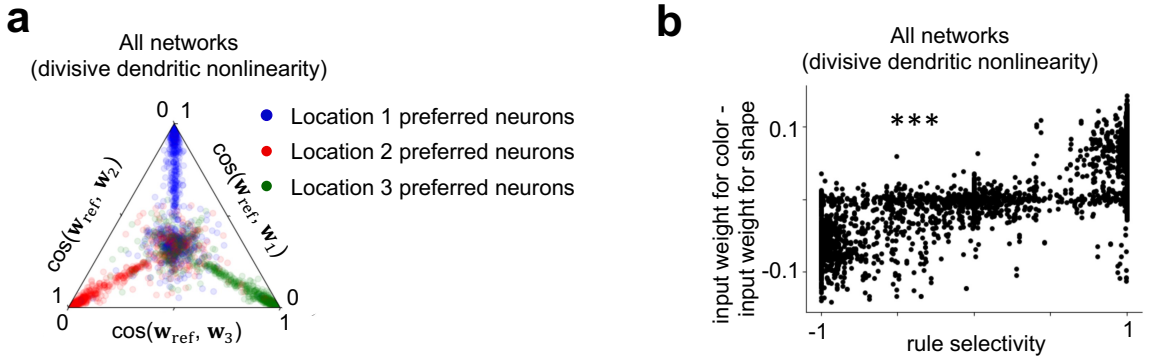
d. Each line represents the mean connection strength onto one SST neuron in the sensorimotor module, from the PFC excitatory neurons that prefer the same rule and the different rule. Bars represent the mean across neurons. PFC excitatory neurons project more strongly to sensorimotor SST neurons that prefer the same rule (c.f. Figure 4d).

e. Each line represents the mean connection strength onto one excitatory neuron of the sensorimotor module, from the sensorimotor SST neurons that prefer the same rule and the different rule. Bars represent mean across neurons. Sensorimotor SST neurons project more strongly to sensorimotor excitatory neurons that prefer the opposite rule (c.f. Figure 4e).

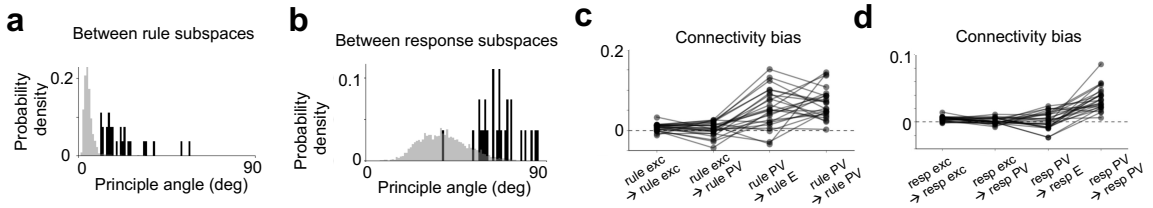
f. Each line represents the mean connection strength onto one excitatory neuron of the sensorimotor module, from the sensorimotor PV neurons that prefer the same rule and the different rule. Bars represent mean across neurons. Sensorimotor PV neurons project more strongly to sensorimotor excitatory neurons that prefer the opposite rule (c.f. Figure 4f).

g - l. Same as **a - f** for networks with the divisive dendritic nonlinearity.

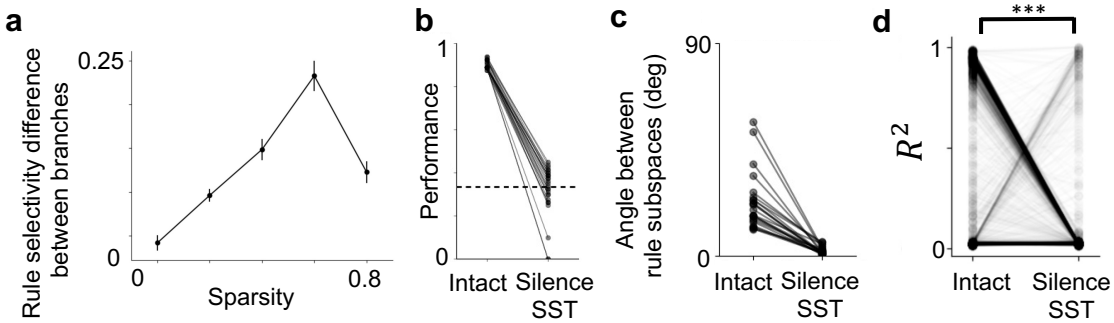
$p < 0.001$ for all panels, Student's t test.



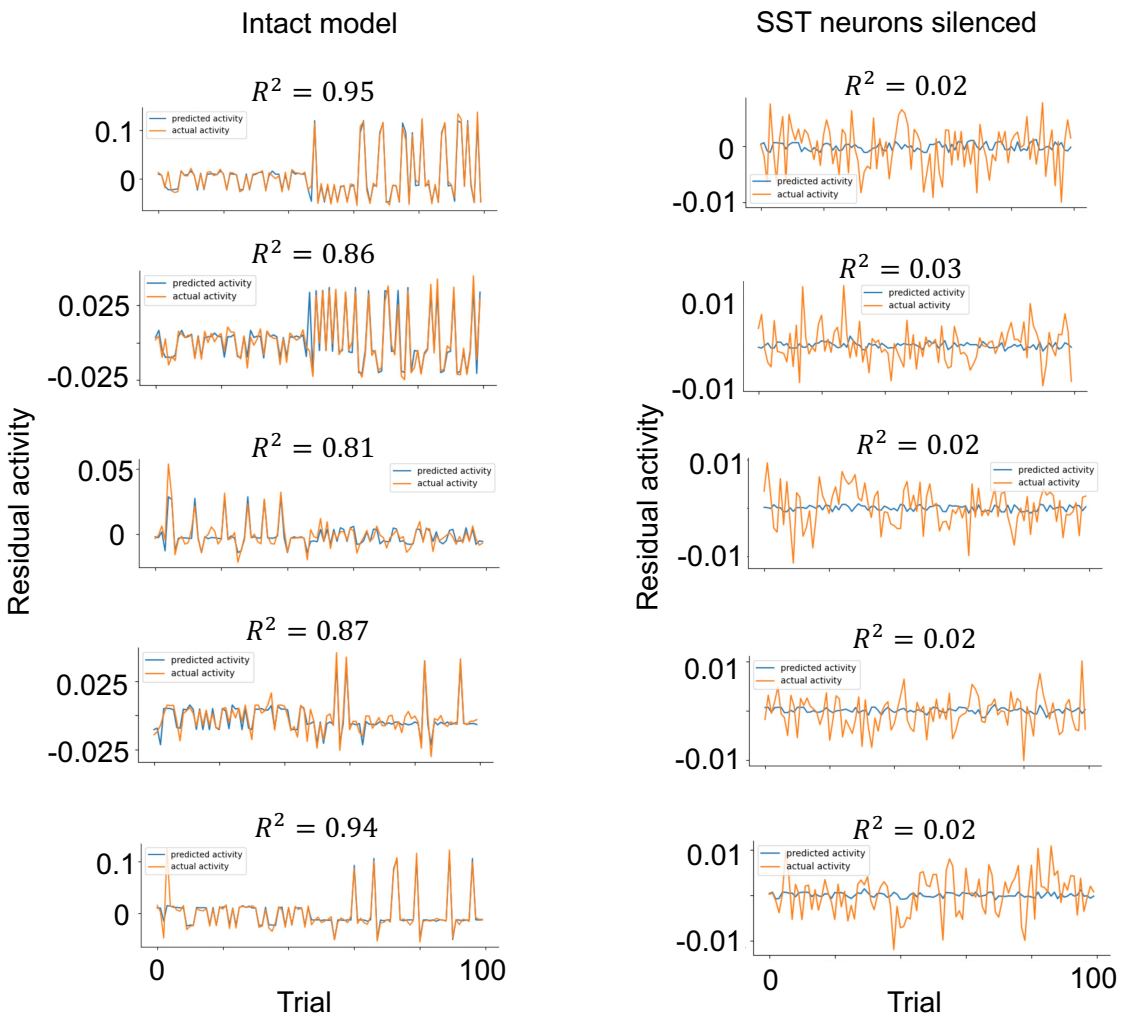
Supplementary Figure 6: The structure in the input weight for networks with divisive dendritic nonlinearity (c.f. Figure 5e, g). For **b**, Pearson's $R = 0.74$, $p < .001$.



Supplementary Figure 7: **a-b**. The principal angle between different rule and response subspaces for networks trained with divisive dendritic nonlinearity (c.f. Figure 6c-d). **c-d**. The connectivity bias between different subpopulations of excitatory and PV neurons in the sensorimotor module of networks with divisive dendritic nonlinearity (c.f. Figure 6g-h).



Supplementary Figure 8: **a**. The relationship between the sparsity of the connection from SST to the dendrites of excitatory neurons in the sensorimotor module, for networks with divisive dendritic nonlinearity (c.f. Figure 7d). **b-d**. The performance (**b**), principal angle between rule subspaces (**c**) and the strength of conjunctive coding (**d**) decreased after silencing SST neurons in the sensorimotor module, for networks trained with divisive dendritic nonlinearity (for **d**, Student's t-test, $p < .001$. Each line represents one neuron. Results are aggregated across networks). c.f. Figure 7e-g.



Supplementary Figure 9: Example neurons in the sensorimotor module showing decreased conjunctive coding of rule and stimulus when SST neurons in the sensorimotor module are silenced. The strength of conjunctive coding is assessed by the R^2 value of a linear regression model where the independent variables are conjunctions of rule and stimulus and the dependent variable is the trial-to-trial residual neural activity unexplained by a model with only rule and stimulus as independent variables (see Methods for details).