Contributions and synaptic basis of diverse cortical neuron responses to task performance

Robert Froemke (robert.froemke@med.nyu.edu)
New York University Grossman School of Medicine https://orcid.org/0000-0002-1230-6811
Michele Insanally
University of Pittsburgh
Badr Albanna
University of Pittsburgh
Jack Toth
University of Pittsburgh
Brian DePasquale
Princeton University
Saba Fadaei
NYU Grossman School of Medicine
Trisha Gupta
University of Pittsburgh
Kishore Kuchibhotla
Johns Hopkins University https://orcid.org/0000-0002-2344-2595
Kanaka Rajan
Icahn School of Medicine at Mount Sinai https://orcid.org/0000-0003-2749-2917

Biological Sciences - Article

Keywords:

Posted Date: May 18th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1628084/v1

Contributions and synaptic basis of diverse cortical neuron responses to task performance

Authors: Michele N. Insanally[1]†*, Badr F. Albanna[2]*, Jack Toth[1], Brian DePasquale[3], Saba Shokat Fadaei[4], Trisha Gupta[1], Kishore Kuchibhotla[5], Kanaka Rajan[6], and Robert C. Froemke[4,7]†

Affiliations:

[1] Departments of Otolaryngology, Neurobiology, University of Pittsburgh School of Medicine, Pittsburgh, PA 15213

[2] Department of Neuroscience, University of Pittsburgh, Pittsburgh, PA 15213

[3] Princeton Neuroscience Institute, Princeton University, Princeton NJ, USA

[4] Skirball Institute for Biomolecular Medicine, Neuroscience Institute, Departments of Otolaryngology, Neuroscience and Physiology, New York University Grossman School of Medicine, New York, NY, 10016, USA

[5] Departments of Psychological and Brain Sciences, Neuroscience, and Biomedical Engineering, Johns Hopkins University, Baltimore, MD, 21218, USA

[6] Department of Neuroscience and Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, New York, NY, 10029, USA

[7] Center for Neural Science, New York University, New York, NY, 10003, USA

* Co-first-authors.

† Co-corresponding authors whom correspondence should be addressed:
Phone: 412-648-4620
Email: mni@pitt.edu

Phone: 212-263-4082
Email: robert.froemke@med.nyu.edu

Abstract

Neuronal responses during behavior are diverse, ranging from highly reliable 'classical' responses to irregular or seemingly-random 'non-classically responsive' firing. While a continuum of response properties is frequently observed across neural systems, little is known about the synaptic origins and contributions of diverse response profiles to network function, perception, and behavior. Here we use a task-performing, spiking recurrent neural network model incorporating spike-timing-dependent plasticity that captures heterogeneous responses measured from auditory cortex of behaving rodents. Classically responsive and non-classically responsive model units contributed to task performance via output and recurrent connections, respectively. Excitatory and inhibitory plasticity independently shaped spiking responses and task performance. Local patterns of synaptic inputs predicted spiking response properties of network units as well as the responses of auditory cortical neurons from *in vivo* whole-cell recordings during behavior. Thus a diversity of neural response profiles emerges from synaptic plasticity rules with distinctly important functions for network performance.

Introduction

Neuronal spiking patterns and responses to sensory input can be remarkably diverse, ranging from completely silent or firing a single action potential to prolonged burst firing or complex sequence generation. Various spiking patterns have been documented throughout brain regions in response to different sensory inputs, in relation to decision making, motor actions, or other task-related signals. The extent of spiking and receptive field heterogeneity is vast, with numerous types of neuronal responses found in many brain areas including visual cortex[1–4], auditory cortex[5–11], somatosensory cortex[12], parietal cortex[3,13,14], frontal cortex[14–18], hypothalamus[19–21], hippocampus[22–24], and the ventral tegmental area[25,26] correlated with sensory, motor, choice, and other task-related signals[1,3,10,13–15,27,28]. These neurons are often described as tuned, untuned, classically responsive, non-classically responsive, mixed selective, or category-free[13,15,17,29].

Recently we showed that classically responsive neurons (e.g., pure tone frequency tuning in auditory cortex) and non-classically responsive neurons (i.e., nominally "non-responsive" neurons) both contained significant information about sensory stimuli and behavioral decisions. This finding suggests that non-classically responsive cells play important yet generally underappreciated roles in perception and behavior[15]. This work is consistent with other recent findings demonstrating that neurons in rat primary visual and parietal cortex can encode sensory and non-sensory factors related to movement, reward history, and decision making[3]. Similarly, a recent study on sequential memory in humans found that both strongly-tuned and weakly-tuned neurons recorded from the medial temporal lobe participated in theta-phase-locked encoding of sequence stimuli[30]. A previous study on working memory in primate prefrontal cortex also revealed that non-selective neurons can contribute to optimal ensemble encoding[18], consistent with our own finding that mixed ensembles of classically and non-classically responsive cells improved

encoding of task variables[15]. Studies of deep neural networks trained to perform a visual recognition task showed that regularizing networks to increase the fraction of 'non-selective' units improved network performance relative to those with greater numbers of 'selective' units[31]. These findings suggest that a diversity of neuronal response types (including neurons nominally thought to be non-responsive) may be a general property of neural networks, and that heterogeneity may be a key feature of the circuit dynamics important for network performance and behavior.

Here we now examine the synaptic basis for various types of spiking response profiles, and how synaptic plasticity learning rules were important for shaping synaptic inputs for spike output and network performance. We leveraged cell-attached, extracellular, and whole-cell recordings from behaving animals alongside recurrent network modeling to explore the synaptic origins and functional contribution of heterogeneous response profiles. Recent advancements using recurrent neural networks (RNN) with spiking units have shown that experimentally-derived synaptic plasticity mechanisms can support stable neuronal assemblies[32] and coordinate memory formation and retrieval[33]. Related work has also shown that spiking RNNs can be trained to perform tasks using general-purpose methods similar to those employed in rate-based networks such as first-order reduced and controlled error (FORCE) training[34,35] but these methods have only been employed as perturbation of networks with static synaptic weights. We combined FORCE training with a dynamic network to create a novel RNN with spiking units and multiple spike-timing-dependent plasticity (STDP) rules to solve a stimulus classification task similar to that of trained rats and mice. Our goal was to determine whether and how classically and non-classically responsive units contribute to task performance, how local synaptic structure (e.g. monosynaptic and disynaptic connections) constrains single-unit response profiles, and if the relationships observed for units in our network model could be applied to neurons *in vivo* during behavior.

Results

Diverse cortical responses measured during behavior in freely-moving rats and head-fixed mice

We recorded from the rodent auditory cortex as animals performed a task requiring them to classify specific tone frequencies. We first trained rats to perform a go/no-go auditory frequency recognition task (**Fig. 1a**) requiring them to behaviorally respond with a nosepoke to a single target tone (4 kHz) for food reward and to withhold responses on non-target tones (0.5, 1, 2, 8, 16, 32 kHz). Rats learned this task within a few weeks of training performing with high d' values (**Fig. 1b**; d' = 2.8±0.1, $p<10^{-4}$, N = 15, Wilcoxon two-sided test). We previously showed that the auditory cortex is required for this task[6,15]. After rats reached behavioral criteria (percent correct: ≥70%, d': ≥1.5), tetrodes were implanted in the right auditory cortex and we recorded from populations of single-units in non-head-fixed animals as they performed this go/no-go task (**Fig. 1c**). Single-unit responses were quite diverse across the population, spanning a range of response types from 'classically responsive' cells that were highly modulated relative to pre-trial baselines during the task to 'non-classically responsive' cells with relatively unmodulated firing rates throughout task performance including cue presentation and behavioral choice.

To capture the continuum of response types, we calculated a 'firing rate modulation index' comparing neural responses during the stimulus and choice periods to baseline values where either positive or negative changes in spike number increase the modulation index (in units of spikes per second). A low value of firing rate modulation index (near 0 spikes/s) corresponds to neurons that were unmodulated relative to baseline ('non-classically responsive') and larger values (≥2 spikes/s) correspond to neurons that were highly modulated ('classically responsive'). The modulation was calculated from the firing rates during the stimulus and choice periods ('stimulus/choice FR') and baseline firing rate ('baseline FR') as

$$\text{modulation} = \sqrt{(\text{stimulus FR} - \text{baseline FR})^2 + (\text{choice FR} - \text{baseline FR})^2}. \quad (1)$$

Most single-units recorded during behavior were generally non-classically responsive, with a minority of cells having more 'classical' responses, e.g., to tone presentation (**Fig. 1d,e**; median firing rate modulation = 0.78 spikes/s, interquartile range = 0.47 – 1.50 spikes/s).

We observed a similar range of neuronal response profiles in cell-attached recordings from the auditory cortex of head-fixed mice trained on an analogous go/no-go auditory frequency recognition task (**Fig. 1f**). Mice were trained to respond to a single target tone (11.2 kHz) by licking for a water reward, and to withhold their response to a single non-target tone (5.6 kHz). Mice learned this task within a few weeks performing at high d' values (**Fig. 1g**; d' = 2.5±0.1, p = 0.016, N = 7 mice, Wilcoxon two-sided test). After reaching behavioral criteria (percent correct ≥70% and d': ≥1.5), animals were implanted with a cranial window that included a small hole for pipette access, and cell-attached recordings were made to measure spike firing during behavior (**Fig. 1h**). We found that neuronal responses in mice were also heterogeneous, including non-classically responsive cells with low firing rate modulation as well as classically-responsive cells that were highly modulated during the task (**Fig. 1i,j**; median firing rate modulation = 2.26, interquartile range = 1.73 – 3.61 spikes/s). These data then led us to wonder how local patterns of single-neuron excitatory and inhibitory inputs related to spike firing and behavioral performance (**Fig. 1k**).

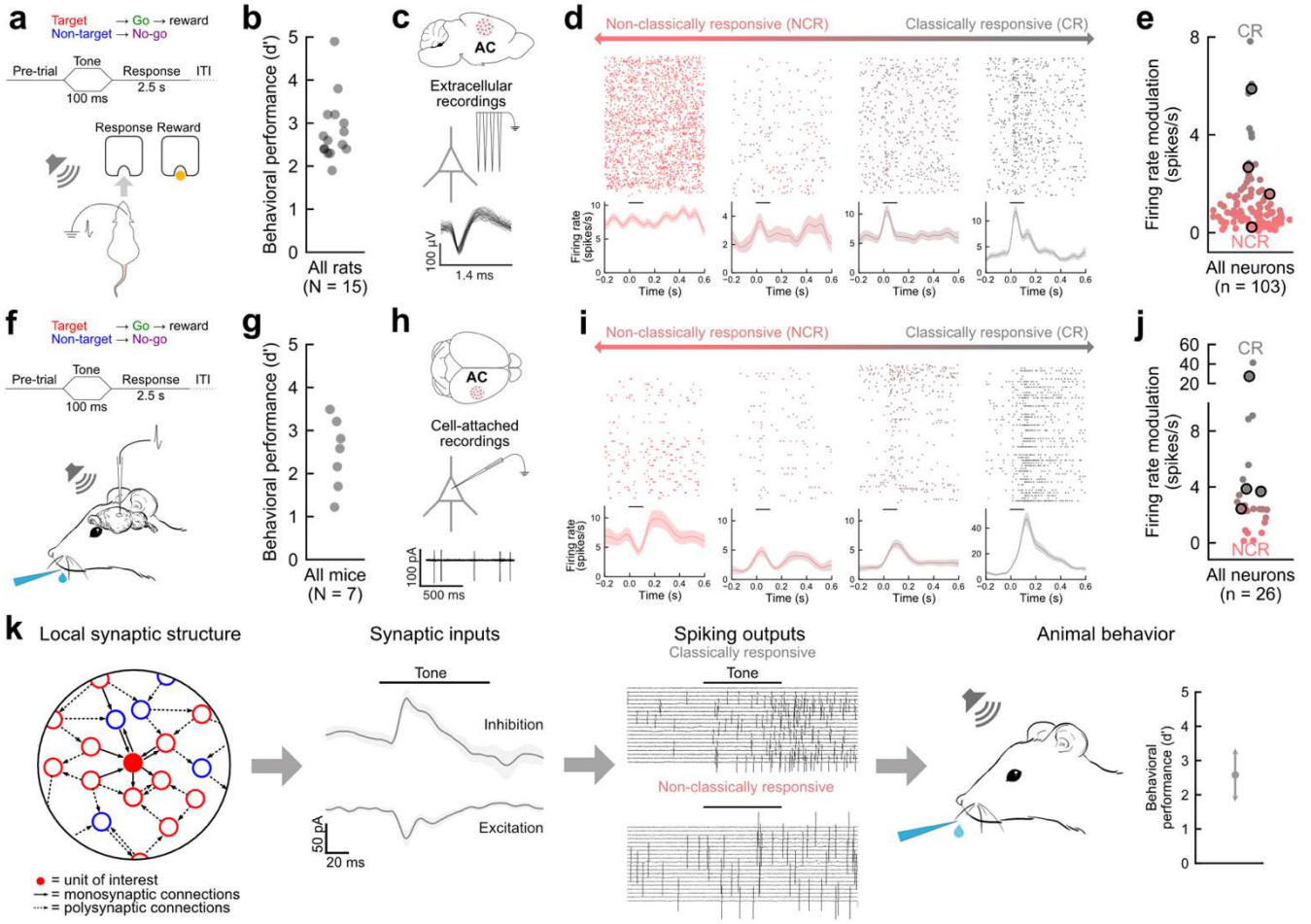

Figure 1. Diverse single-unit responses measured in rodent auditory cortex during behavior. a, Schematic of behavior and extracellular tetrode recordings from auditory cortex of rat performing frequency recognition go/no-go task. **b**, Asymptotic behavioral performance for all rats (d' = 2.8±0.1, $p<10^{-4}$, N = 15, Wilcoxon two-sided test). **c**, Example single-unit recording from rat auditory cortex during behavior. **d**, Rasters and peri-stimulus time histograms (PSTHs) for four cortical neurons exemplifying the range from non-classically responsive (red, NCR) to classically responsive (gray, CR). Lines in PSTH, mean firing rate; shading, S.E.M. Horizontal bar, tone duration. **e**, Summary of firing rate modulation for all cortical neurons recorded during behavior (n=103). Outlined circles, units from **d**. Median firing rate modulation = 0.78 spikes/s (inter-quartile range 0.47 – 1.50 spikes/s). **f**, Cell-attached recordings from auditory cortex of mouse performing frequency recognition go/no-go task. **g**, Asymptotic behavioral performance for all mice (d' = 2.45±0.11, p = 0.016, N = 7 mice, Wilcoxon two-sided test). **h**, Example cell-attached recording from mouse auditory cortex during behavior. **i**, Rasters and PSTHs for four example recordings. **j**, Firing rate modulation for all cell-attached recordings (n=26) from mouse auditory cortex during behavior (median firing rate modulation = 2.26, interquartile range = 1.73 - 3.61 spikes/s). **k**, Diagram of relationship between local synaptic structure, synaptic inputs, spiking outputs, and behavior.

A spiking RNN model incorporating STDP rules captures *in vivo* cortical dynamics

To relate inputs and outputs over the response-type continuum, we developed a spiking RNN model trained to perform a similar go/no-go stimulus classification task as behaving animals (**Fig. 2a-c**; d' = 4.6±0.1, p=0.0078, N = 8 networks, Wilcoxon two-sided test). All networks contained 1000 units (200 inhibitory, 800 excitatory); 200 excitatory units received inputs and the remaining 600 excitatory units projected onto the readout node. The activity of this node is the dynamic signal which represents the response of our network to the incoming stimuli: 'go' is represented by an increase in node activity during the choice period, and 'no-go' is represented by node activity that remains at pre-trial baseline. All units were recurrently connected with a 5% random connection probability. The network was trained to perform the task via a version of FORCE designed for spiking networks[34,35] (i.e., least-squares modification of the output weights with feedback) combined with STDP synaptic plasticity learning rules acting on the recurrent weights (**Fig. 2a, Extended Data Fig. 1a-d**). The network was trained to use the minimal amount of external stimulus input while being able to perform the task to mirror the behavioral errors seen during animal performance (**Extended Data Fig. 1e**).

Our model included biologically-motivated and experimentally-constrained excitatory and inhibitory synaptic STDP. Excitatory-to-excitatory synapses were modified by classic pairwise Hebbian homosynaptic plasticity[36,37], and inhibitory-to-excitatory synapses were modified by a homosynaptic rule which strengthens synapses when units fire synchronously regardless of order[38–41]. Excitatory-to-excitatory and inhibitory-to-excitatory synapses were also adjusted by homeostatic mechanisms of heterosynaptic plasticity which prevented any one presynaptic connection from dominating (heterosynaptic balancing, β) or postsynaptic connection from dropping out (heterosynaptic enhancement, δ). These heterosynaptic changes occurred simultaneously with homosynaptic mechanisms and thus were

qualitatively distinct from other types of homeostatic regulation of input weight distributions such as synaptic scaling.

While FORCE was originally designed to construct networks that can solve tasks without modification of the underlying recurrent synapses, in our network, STDP constrains FORCE to find solutions consistent with biological plasticity rules. FORCE and STDP can operate in parallel because each mechanism can be targeted to a different set of connections in the model: STDP modifies the recurrent synapses while FORCE modifies the connections to the readout node. In this way, STDP directly shaped the inherent recurrent dynamics of our network while FORCE determined how those dynamics were harnessed to perform the task. Since FORCE was originally designed to operate in networks with fixed recurrent weights and in our network these evolve, gross changes induced by STDP must be complete before FORCE can adapt to smaller synaptic changes. To ensure this, STDP was first activated without FORCE to allow for initial major synaptic restructuring to occur before FORCE training began and the two mechanisms continued in parallel (see Methods). The particular mechanisms and parameters chosen for STDP in this model were not optimized for the task a priori, and as such STDP should not have trivially improved performance. In general, we found that by using this procedure FORCE was compatible with STDP over a wide range of STDP parameters (**Extended Data Fig. 1f-h**). All networks without STDP active were trained with FORCE for the same number of trials as those with STDP.

We found that with STDP the spiking responses of individual network units closely approximated the distribution of firing rate modulations observed experimentally in the auditory cortex *in vivo* (**Fig. 2e**; $p = 0.27$, Kolmogorov-Smirnov test). Using a statistical threshold to identify non-classically responsive units as previously described[15] (firing rate change from baseline < 0.2 spikes/s during stimulus and choice

periods, see Methods), we found the relative fractions of classically and non-classically responsive units were also comparable to experimental measurements (**Extended Data Fig. 2a-c**; ~40-50% non-classically responsive, 50-60% classically responsive). Using a single-trial, interspike interval (ISI)-based, Bayesian decoder we recently described[15], we found that task information was encoded in the activity of both classically and non-classically responsive RNN units (**Fig. 2f**).

To assess whether STDP altered response profile distributions, we compared the distribution of units before STDP was applied (pre-STDP) to those after (post-STDP). In pre-STDP networks, the recurrent weights were fixed during FORCE training whereas in post-STDP networks the recurrent weights evolved according to the STDP rules described above. Pre- and post-STDP networks were constructed in pairs with the same set of initial recurrent weights so that pre-STDP networks represent the behavior of the network with FORCE alone before STDP was active. The post-STDP response profile distribution differed substantially from pre-STDP networks, such that post-STDP networks exhibited more non-classically responsive units than pre-STDP networks (**Fig. 2g**; median post-STDP modulation = 1.52 spikes/s vs. pre-STDP modulation = 2.25 spikes/s, $p < 10^{-5}$, Mann-Whitney two-sided U-test with Bonferroni correction). To understand how the synaptic changes induced by STDP resulted in this shift, we created two control networks with engineered weight matrices based on the full post-STDP weights. To determine whether the post-STDP response profile distribution resulted from a shift in mean weight strength, we first created networks with weights generated as in the pre-STDP condition but with mean values matched to those found post-STDP (**Fig. 2g**; 'Mean-match', weights drawn from uniform distribution from 0 to 2x mean, median post-STDP modulation = 1.52 vs. mean-match modulation = 1.46, $p<10^{-4}$, N = 8 networks; **Extended Data Fig. 3a**). Second, we created networks in which inhibitory-to-excitatory and excitatory-to-excitatory synaptic weights were shuffled at random to new pre- and

postsynaptic target units to preserve the full weight distribution created by STDP but remove any synaptic correlations (**Fig. 2g**, 'Shuffle', median post-STDP modulation = 1.52 vs. shuffle modulation = 1.48, p=0.0017, N = 8 networks, **Extended Data Fig. 3a**). Both control conditions were retrained with FORCE after weight matrices were altered. For both the 'mean-match' and 'shuffle' controls, the distributions of responses closely followed the post-STDP distribution; however, the proportion of inactive units (firing rate < 1 spike/trial on average) significantly increased to ~5% of the full network (**Fig. 2h**; comparisons to post-STDP p = 0.001, Mann-Whitney two-sided U test with Bonferroni correction). These results show that the overall synaptic weight changes induced by STDP could account for the increase in the fraction of non-classically responsive units. However, the synaptic correlations induced by STDP were required to facilitate full network engagement (i.e., no inactive units).

To determine how STDP shaped synaptic weight strengths and structure (i.e., network topology), we compared the synaptic weight strengths pre-STDP to post-STDP. Over the course of training, we found that STDP mechanisms made significant modifications to both the excitatory-to-excitatory and inhibitory-to-excitatory synaptic weight distributions and topologies. Median weights shifted relative to pre-STDP initial values for both types of synapses and the distributions of synaptic weights became skewed (**Extended Data Fig. 3b**), as has been observed as has been observed in rat visual cortex[42]. We asked if selective weight rescaling by STDP preserved the random structure of the initial pre-STDP network or created systematic patterns of synaptic connectivity (i.e., "small-world" network topology) as previously described in rat somatosensory cortex[43,44]. We compared the topology of the synaptic weight matrix post- and pre-STDP and found that STDP generated a more clustered network structure consistent with both experimental observations[44] and computational studies[32,33] (**Extended Data Fig. 3c**).

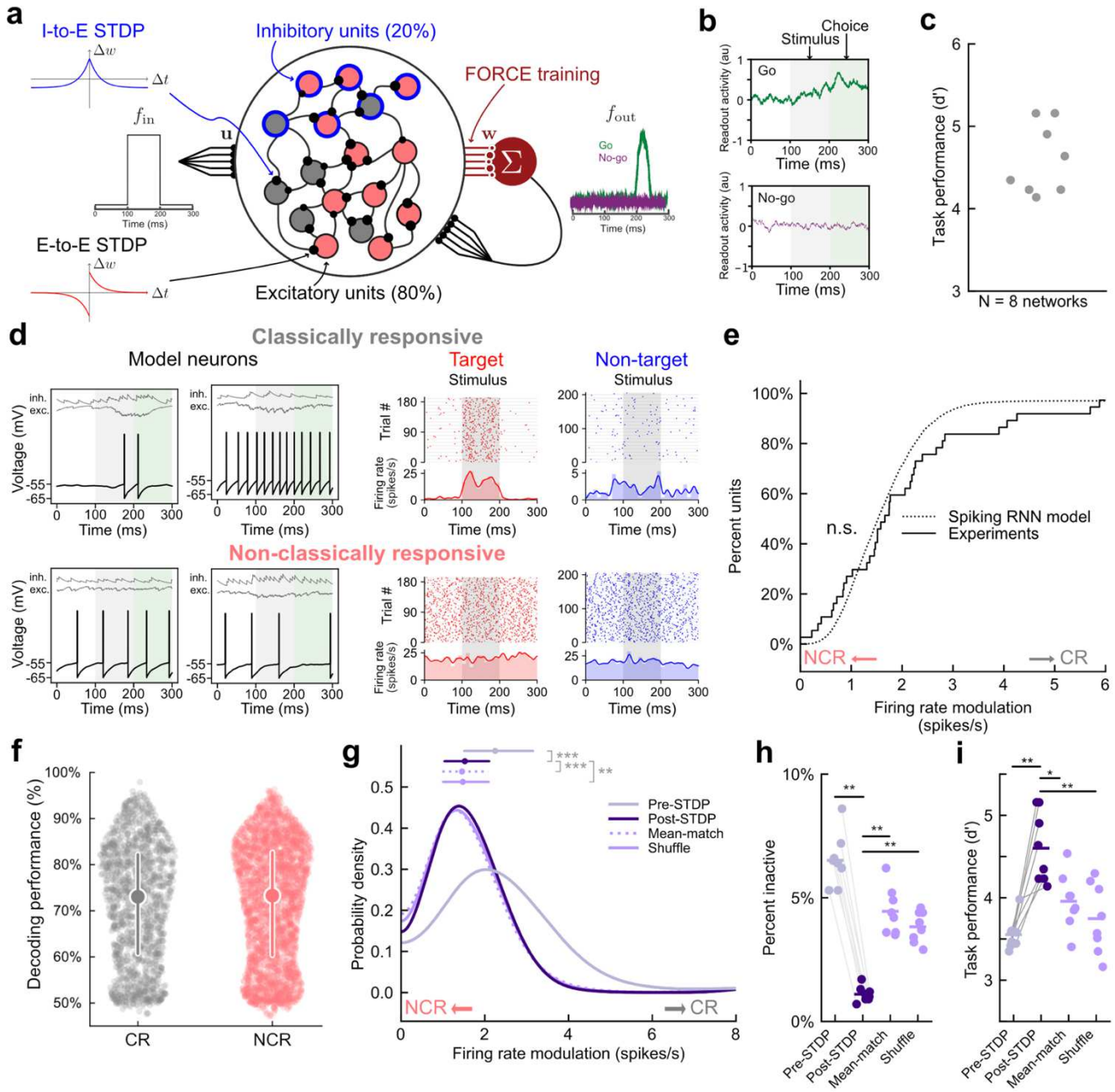

Figure 2. A spiking RNN model incorporating STDP rules recapitulating *in vivo* cortical neuronal dynamics. **a,** Schematic of spiking RNN model trained to complete go/no-go stimulus recognition task. Networks consisted of 80% excitatory and 20% inhibitory units. 25% of excitatory units received direct current as the stimulus (1 target stimulus, 6 non-target stimuli) and remaining 75% were output units which project to the readout node (maroon) and received feedback from readout node. Weights to the readout node trained were via FORCE. Excitatory-to-excitatory and inhibitory-to-excitatory synapses were modified by separate STDP mechanisms. **b,** Example network outputs on 'go' trial (in response to

'target' stimulus) and 'no-go' trial (in response to 'non-target' stimulus). White, pre-trial baseline; gray, stimulus period; green, choice period. **c,** Asymptotic task performance (d' = 4.6±0.1, p=0.0078, N = 8 networks, Wilcoxon two-sided test) **d,** Top left, example voltage traces from two trials of a classically responsive unit. Top right, corresponding spike rasters across trials and PSTHs to target (red) and non-target stimuli (blue). Bottom left, two example voltage traces from a non-classically responsive unit. Bottom right, corresponding spike rasters across trials and PSTHs. **e,** Cumulative distribution of single-unit firing rate modulation for spiking RNN model (dotted) and experimental data (solid). Small values of the firing rate modulation correspond to non-classical response (NCR) profiles; high values correspond to classical response (CR) profiles. Difference between simulated and experimental distributions were not statistically significant (p = 0.27, Kolmogorov-Smirnov test). **f,** Decoding performance for single units in one network (n=1000) for classically responsive units (CR, grey, left) and non-classically responsive units (NCR, red, right). Circle and line represent median and interquartile range respectively. **g,** Probability density of firing rate modulation for individual units. Gray, pre-STDP; purple, post-STDP; dotted light purple, mean-matched control; solid light purple, shuffle control. Summary circles and bars above distributions represent median and interquartile range, respectively (median post-STDP modulation = 1.52 vs. pre-SDTP modulation = 2.25, $p < 10^{-5}$, vs. mean-match modulation = 1.46, $p<10^{-4}$, 52 vs. shuffle modulation = 1.48, p=0.0017, N = 8 networks in all groups, Mann-Whitney two-sided U-test with Bonferroni correction). **h,** Percent of inactive units for pre-STDP, post-STDP, and controls (defined as firing rate <1 spikes/s, all comparisons to post-STDP, p = 0.001, Mann-Whitney two-sided U test with Bonferroni correction). **i,** Task performance for pre-STDP, post-STDP, and controls (Mean shifts relative to pre-STDP, post-STDP: Δd' = 0.97, p = 0.0014, mean-match: Δd' = 0.44, p = 0.04, shuffle: Δd' = 0.15, p = 0.74, N = 8 networks in all groups, Wilcoxon two-sided test with Bonferroni correction).

The modifications to the weight matrix made by STDP improved task performance (**Fig. 2i**; **median** pre-SDTP d' = 3.52 vs. post-STDP d' = 4.49, p = 0.0014, N= 8 networks, Wilcoxon two-sided test with Bonferroni correction). This increase in performance was not observed in either the mean-matched or shuffled control conditions, demonstrating that the detailed connectivity structure of synaptic weights created by STDP were required to improve performance (**Fig. 2i**; mean pre-STDP d' = 3.52 vs. mean-match d' = 3.96, p = 0.04, and shuffle d' = 3.67, p = 0.74, N = 8 networks in all groups, Wilcoxon two-sided test with Bonferroni correction). These results show that STDP shifts the weight matrix into a regime leading to improved task performance while maintaining a consistent level of network unit engagement.

Classically and non-classically responsive RNN units contribute to task performance via distinct mechanisms

Given that STDP increased both the prevalence of non-classically responsive units as well as network task performance, we next explored how classically responsive and non-classically responsive units contributed directly to task performance. We first evaluated the output connections to the readout node and recurrent weights between units to determine whether there were systematic differences between these response profiles. Both classically and non-classically responsive units spanned a similar range of values, however classically responsive units had larger weights onto the readout node than non-classically responsive units (**Fig. 3a**; $p < 10^{-5}$, Levene's test). This suggests that both classes contribute directly to task performance by driving the network output, but classically responsive units may affect performance specifically via their effect on the readout node. In contrast, non-classically responsive units had stronger recurrent projections than classically responsive units to both subpopulations (**Fig. 3b**; $p < 10^{-5}$, for all comparisons between NCR and CR, Mann-Whitney U test two-sided with Bonferroni-correction). This result, coupled with the observation that non-classically responsive units generally had higher firing rates

(**Fig. 3c**; median NCR = 19.1 spikes/s vs. median CR = 16.5 spikes/s, $p < 10^{-5}$, Mann-Whitney two-sided U test) suggests non-classically responsive units may play a privileged role in generating task-related dynamics through their effect on recurrent network activity.

To test this hypothesis, we transiently inactivated classically and/or non-classically responsive units during task performance. Inactivation of non-classically responsive units targeted the least modulated units in the network first while inactivation of classically responsive units targeted the most modulated units. During inactivation, we silenced output connections and replaced recurrent activity with average-firing-rate-matched Poisson noise to control for changes in the overall level of recurrent synaptic current. Completely inactivating either subpopulation impaired task performance suggesting that both subpopulations are important for network dynamics. Inactivating a relatively small number of highly non-classically responsive units (60 units, top 10% most non-classically responsive) had a significantly larger effect on performance than inactivating classically responsive units (**Fig. 3d, Extended Data Fig. 4a**; inactivating 10% most non-classically responsive $\Delta d' = -1.6$ vs. inactivating 10% most classically responsive $\Delta d' = -0.9$, $p < 10^{-4}$, Mann-Whitney two-sided U test with Bonferroni-correction). As the number of inactivated units increased, however, task performance continued to degrade and eventually the effects of inactivating 300 units (i.e., 50%) of units from either subpopulation were comparable (**Fig. 3d**; inactivating 50% most non-classically responsive $\Delta d' = -2.3$ vs. inactivating 50% most classically responsive $\Delta d' = -2.6$, $p = 0.06$, Mann-Whitney two-sided U test with Bonferroni correction).

The types of errors produced ('false alarms' on non-target trials and 'misses' on target trials) did not differ when inactivating either classically or non-classically responsive units (**Fig. 3e**; false alarms vs. misses inactivating 50% most classically responsive, $p = 0.26$, inactivating 50% most non-classically responsive,

p = 0.21, Mann-Whitney two-sided U test with Bonferroni correction). However, we observed a qualitative difference in the network dynamics that produced these errors depending on which subpopulation was inactivated. Inactivating classically responsive units let to greater variability in the readout node activity during the choice period (**Fig**. **3f;** root mean squared error during response period at 50% inactivation of most classically responsive = 0.48, 50% most non-classically responsive = 0.44, $p < 10^{-5}$, Mann-Whitney two-sided U test with Bonferroni correction). In contrast, inactivating non-classically responsive units caused a greater shift in readout node baseline activity (**Fig**. **3g;** mean shift in baseline activity at 50% inactivation of most classically responsive = -0.25 vs. 50% most non-classically responsive = 0.47, $p < 10^{-5}$, Mann-Whitney two-sided U test with Bonferroni correction). This suggests that while both response types are essential for proper network function, non-classically responsive units served to set boundaries which constrain overall network dynamics to the task-relevant subspace (i.e., readout node activity close to 0 except during 'go' responses) whereas classically responsive units affected dynamics within those boundaries.

We asked if these impairments resulted from silencing output connections or interfering with recurrent activity by either selectively inactivating output connections or recurrent connectivity. Inactivation of the output connections alone in either subpopulation impaired performance. Removing output connections from 10% of non-classically responsive units led to a significant decrease in task performance (**Extended Data Fig. 4b**; removing output connections from 10% most non-classically responsive $\Delta d' = -1.2$ vs. 10% most classically responsive $\Delta d' = -0.6$, $p < 10^{-5}$, Mann-Whitney two-sided U test with Bonferroni correction). For greater than 20% inactivation, silencing output connections from classically responsive units resulted in a stronger decrease in performance (**Extended Data Fig. 4b**; $p < 0.05$, Mann-Whitney two-sided U test with Bonferroni correction). This indicates that while classically responsive units

contributed more to task performance via their output connections overall, a small fraction of highly non-classically responsive units were also critical for high levels of performance via their output projections. In contrast to silencing output connections, disabling the recurrent contributions of non-classically responsive units (by replacing units with firing-rate-matched Poisson noise) resulted in a larger impairment in performance than classically responsive units for all numbers of units inactivated (**Extended Data Fig. 4c**; disabling 10% most non-classically responsive output units $\Delta d' = -0.7$ vs. 10% most classically responsive units $\Delta d' = -0.4$, $p=0.016$, Mann-Whitney U two-sided with Bonferroni-correction). Taken together, these selective inactivation experiments show that classically responsive neurons contribute more to task performance through their output projections while non-classically responsive neurons contribute primarily through their recurrent activity.

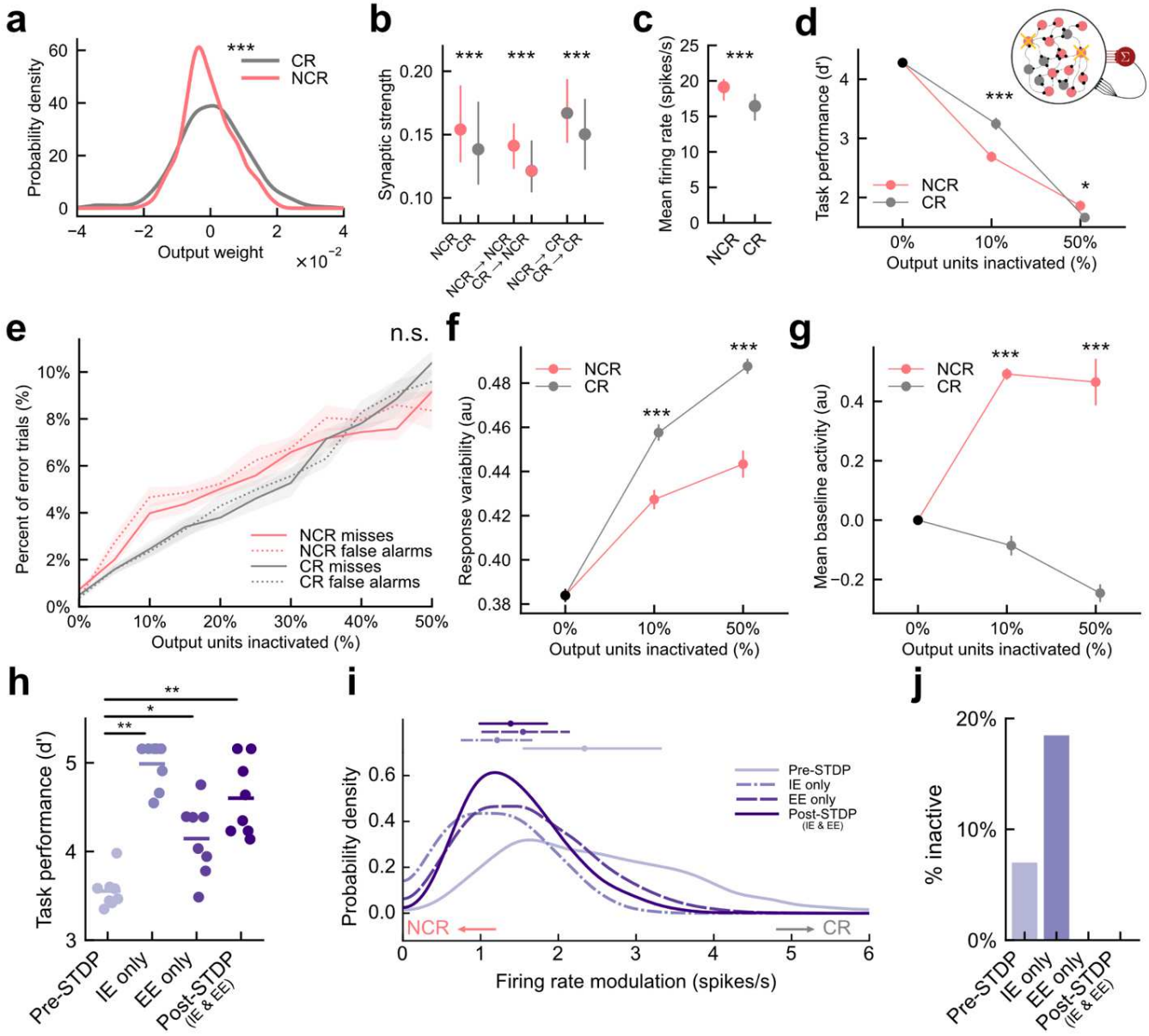

Figure 3. Classically and non-classically responsive units contribute to task performance and these response profiles shaped by excitatory and inhibitory STDP. a, Probability density function for output weights from non-classically responsive units (red, NCR) and classically responsive units (grey, CR). N=8 networks, n=4,800 units, $p<10^{-5}$, Levene's test. **b,** Outgoing recurrent synaptic weights from non-classically responsive units (left) and classically responsive units (right). Circles and lines represent median and interquartile range, respectively. Synaptic weights from NCRs were greater overall and when conditioned on target subpopulation (NCR, CR). $p < 10^{-5}$ for all comparisons between NCR and CR, Mann-Whitney U test two-sided with Bonferonni-correction. **c,** Average firing rates of non-classically responsive units (red, NCR) were higher than those of classically responsive cells (grey, CR). Circles and

lines represent median and interquartile range, respectively. Median NCR = 19.1 spikes/s vs. median CR = 16.5 spikes/s, $p < 10^{-5}$, Mann-Whitney two-sided U test **d,** Task performance as a function of output units inactivated for non-classically responsive units only (red, NCR), classically responsive units only (grey, CR). Points and bars represent mean and S.E.M., respectively. Inactivating 60 units i.e. 10% most non-classically responsive $\Delta d' = -1.58$ vs. inactivating 10% most classically responsive $\Delta d' = -0.92$, $p < 10^{-4}$; inactivating 300 units i.e. 50% most non-classically responsive $\Delta d' = -2.27$ vs. inactivating 50% most classically responsive $\Delta d' = -2.61$, $p = 0.061$, Mann-Whitney two-sided U test with Bonferroni correction. **e,** Error rates for misses (solid) and false alarms (dotted) as a function of output units inactivated for non-classically responsive units only (red, NCR) and classically responsive units only (grey, CR). Lines and shaded areas represent means and S.E.M. False alarms vs. misses inactivating 50% most classically responsive, $p = 0.26$, inactivating 50% most non-classically responsive, $p = 0.21$, Mann-Whitney two-sided U test with Bonferroni correction. **f,** mean squared error for readout node activity during choice period as a function of output units inactivated for non-classically responsive units only (red, NCR) and classically responsive units only (grey, CR). Dots and vertical lines represent means and S.E.M.. At 10% and 50% inactivation, $p < 10^{-5}$, Mann-Whitney two-sided U test with Bonferroni correction **g,** mean readout node activity during baseline pre-stimulus as a function of output units inactivated for non-classically responsive units only (red, NCR) and classically responsive units only (grey, CR). At 10% and 50% inactivation, $p < 10^{-5}$, Mann-Whitney two-sided U test with Bonferroni correction. **h,** Asymptotic task performance for networks without STDP (pre-STDP), with only inhibitory-to-excitatory STDP (IE only), with only excitatory-to-excitatory STDP (EE only), or all STDP rules (post-STDP). N = 8 matched networks per group each seeded with identical initial weights. Circles represent individual networks and bars represent means. Increase in performance relative to pre-STDP with IE only, $\Delta d' = 1.43 \pm 0.09$, $p = 0.0014$, EE only STDP: $\Delta d' = 0.59 \pm 0.18$, $p = 0.011$, both forms of STDP: $\Delta d' = 1.05 \pm 0.19$, $p = 0.0014$, Wilcoxon two-sided test with Bonferroni correction). **i,** Probability density of firing rate modulation for individual output units for networks without STDP (pre-STDP, light purple solid), with only inhibitory-to-excitatory STDP (IE only, light purple dot dashed), with only excitatory-to-excitatory STDP (EE only, purple dashed), or all STDP rules (post-STDP, purple). Summary circles and bars above distributions represent median and interquartile range, respectively. Median shift in firing rate modulation relative to pre-STDP, Post-STDP: Δmodulation = -0.95 spikes/s, IE only: Δmodulation = -1.13 spikes/s**,** EE only: Δmodulation = -0.79 spikes/s, $p < 10^{-5}$ for all comparisons, Mann-Whitney U test two-sided Bonferroni-correction. Inactive units excluded and percentage shown in **j**.

Effect of network parameters on ensemble diversity

Three of the main parameters in the network model are network size, fraction of inhibitory units, and connection probability. We systematically varied each of those parameters (leaving other parameters fixed) to assess their effect on the distribution of classically responsive and non-classically responsive network unit response profiles. Increasing network size while keeping the average number of connections per unit constant increased the proportion of classically responsive units (**Extended Data Fig. 5a, left**; $p < 10^{-5}$ for 2000 units vs 1000 and 500, Mann-Whitney two-sided U test with Bonferroni correction) which was primarily a result of higher variability in the choice-related responses of network units (**Extended Data Fig. 5a, middle, right**; $p < 10^{-5}$ all comparisons, Levene's test with Bonferroni-correction). Increasing the fraction of inhibitory units reduced the overall responsiveness (**Extended Data Fig. 5b, left**; $p < 10^{-5}$ all comparisons, Mann-Whitney two-sided U test with Bonferroni correction) by decreasing the magnitude of both stimulus and choice-related activity (**Extended Data Fig. 5b, middle, right**; stimulus, $p < 10^{-5}$ all comparisons, Mann-Whitney two-sided U test with Bonferroni correction; choice, $p < 10^{-5}$ all comparisons, Levene's test with Bonferroni-correction). Finally, increasing the connection density of the network increased responsiveness (**Extended Data Fig. 5c, left**; $p < 10^{-5}$ all comparisons, Mann-Whitney two-sided U test with Bonferroni correction) despite the fact that connection weights were scaled down as connection probability increased. This increase was driven by an increase in stimulus-related activity (**Extended Data Fig. 5c, middle, right**; $p < 10^{-5}$ all comparisons**,** Mann-Whitney two-sided U test with Bonferroni correction).

Synaptic mechanisms shape response profile distributions and task performance

Our exploration of network parameters indicated that by changing connectivity statistics – perhaps simulating the effects of synaptic plasticity mechanisms – networks can adjust the relative fraction of

classically responsive and non-classically responsive units. Therefore, we next examined the sensitivity of response profiles to the details of STDP in the model. To further understand how STDP improved task performance and determine the specific role of inhibitory and excitatory STDP mechanisms, we selectively included either only excitatory-to-excitatory or inhibitory-to-excitatory plasticity mechanisms during training and evaluated the effect on task performance and response profile distributions.

Selectively enabling either only inhibitory-to-excitatory or excitatory-to-excitatory plasticity boosted task performance relative to pre-STDP networks. However, including only inhibitory-to-excitatory plasticity produced performance gains comparable to the post-STDP condition where both rules were active (**Fig. 3h**; increase in task performance relative to pre-STDP with only inhibitory-to-excitatory STDP, $\Delta d' = 1.4 \pm 0.1$, $p = 0.0014$, increase in performance with only excitatory-to-excitatory STDP: $\Delta d' = 0.6 \pm 0.2$, $p = 0.011$, increase in performance with both forms of STDP: $\Delta d' = 1.1 \pm 0.2$, $p = 0.0014$, Wilcoxon two-sided test with Bonferroni correction). Plasticity in excitatory-to-excitatory and inhibitory-to-excitatory synapses were both sufficient to shift the response profile distribution towards non-classically responsive activity and create a distribution similar to that observed in the full post-STDP model (**Fig. 3i**; median shift in firing rate modulation relative to pre-STDP, Post-STDP: Δmodulation = -0.95 spikes/s, IE only: Δmodulation = -1.13 spikes/s**,** EE only: Δmodulation = -0.79 spikes/s, $p < 10^{-5}$ for all comparisons, Mann-Whitney U test two-sided Bonferroni-correction). Notably, including inhibitory-to-excitatory plasticity alone produced a larger number of inactive units (**Fig. 3j**, 'IE only'; firing rate < 1 spike/trial on average). While inhibitory-to-excitatory synaptic plasticity was sufficient to improve performance to post-STDP levels, it decoupled a large fraction of network units in the process; however, in tandem with excitatory-to-excitatory plasticity performance gains can occur while all units remained engaged during task performance.

These plasticity mechanisms differed in how they shifted response profiles into a non-classically responsive regime. While both inhibitory-to-excitatory and excitatory-to-excitatory plasticity reduced firing rate modulation during the stimulus period (**Extended Data Fig. 6a, left**; median shift relative to pre-STDP, post-STDP: Δstimulus = -0.52 spikes/s**,** IE only: Δstimulus = -0.70 spikes/s, EE only: Δstimulus = -0.67 spikes/s, $p < 10^{-5}$ for all comparisons to pre-STDP, Mann-Whitney U test two-sided Bonferroni-correction), only excitatory-to-excitatory plasticity shifted choice modulation towards post-STDP values (**Extended Data Fig. 6a, right**; median shift relative to pre-STDP, post-STDP: Δchoice = -0.95 spikes/s, $p < 10^{-5}$, IE only: Δchoice = -0.07 spikes/s, $p = 0.078$, EE only: Δchoice = -0.97 spikes/s, $p<10^{-5}$, Mann-Whitney U test two-sided Bonferroni-correction). Furthermore, the range of both stimulus and choice related modulation values decreased when inhibitory-to-excitatory mechanisms were included regardless of whether excitatory-to-excitatory mechanisms were present (**Extended Data Fig. 6a;** IE only vs. Pre-STDP and Post-STDP vs. EE only, $p<10^{-3}$, Levene's test with Bonferroni-correction). This suggests that excitatory-to-excitatory plasticity shifts response profiles toward non-classically responsive activity by shifting median responses closer to zero modulation while inhibitory-to-excitatory plasticity also constrains the range of modulation observed during each trial period.

We next sought to understand how the three forms of plasticity (homosynaptic STDP, heterosynaptic balancing β, and heterosynaptic enhancement δ) adjusted the population of unitary response profiles. To do this, we selectively increased or decreased the strength of each type of plasticity relative to default values for either excitatory-to-excitatory synapses or inhibitory-to-excitatory synapses. Altering the strength of excitatory STDP did not change the response profile distribution (**Extended Data Fig. 6b, left**); however, excitatory heterosynaptic mechanisms significantly modified response properties

(**Extended Data Fig. 6b, center, right**). Strengthening excitatory heterosynaptic balancing (β) increased the firing rate modulation of network units (**Extended Data Fig. 6b**, **center**; $p < 10^{-5}$ for 2x vs 0.5x strength, Kolmogorov-Smirnov test with Bonferroni-correction), whereas strengthening heterosynaptic enhancement (δ) decreased modulation (**Extended Data Fig. 6b**, **right**; $p < 10^{-5}$ for 2x to 0.5x strength, Kolmogorov-Smirnov test Bonferroni-correction). For inhibitory plasticity, heterosynaptic enhancement (δ) increased the modulation of network units in opposition to excitatory-to-excitatory heterosynaptic enhancement (**Extended Data Fig. 6c, right**; $p < 10^{-5}$ for 2x vs 0.5x strength, Kolmogorov-Smirnov test with Bonferroni-correction). Similarly, inhibitory homosynaptic terms also increased the strength of classical responses (**Extended Data Fig. 6c, left**; $p < 10^{-5}$ for 2x vs 0.5x strength, Kolmogorov-Smirnov test with Bonferroni-correction). Strengthening inhibitory heterosynaptic balancing, β, had the effect of expanding the dynamic range of response profiles without shifting the median, resulting in a greater diversity of response types (**Extended Data Fig. 6c, center**; $p < 10^{-5}$ for 2x vs 0.5x strength, Kolmogorov-Smirnov test with Bonferroni-correction). These results suggest that while excitatory-to-excitatory homosynaptic plasticity has minimal effect on response types, each other synaptic mechanism provides complementary constraints on the range and median of the response profile distribution.

Specific local synaptic patterns predict response properties of diverse units

As STDP of excitation and inhibition shaped response profiles across the network, we next determined how individual classically and non-classically responsive units were embedded in the network by examining their synaptic input and output patterns. We analyzed the responses of the group of excitatory units which projected to the readout node since these neurons were not trivially stimulus responsive (via direct stimulus inputs) and displayed the greatest range of stimulus and choice modulation (‘output units’; **Extended Data Fig. 2c**, **center**).

Although the strengths of direct (i.e., monosynaptic) inputs onto a given unit should predominantly determine the response profiles of individual units, higher order correlations may also be relevant for determining single-unit responses. To assess the contributions of direct and higher-order synaptic connections (e.g., disynaptic, trisynaptic, etc.), we adapted a recent approach[45] to systematically decompose the weight matrix into local patterns of connectivity or "motifs" (**Fig. 4a**), starting with small numbers of synaptic connections (lower-order motifs) and progressing to motifs with larger numbers of connections (higher-order motifs). These local patterns have been shown to predict a variety of network phenomena including cross-correlations and the dimensionality of network dynamics[46]. Motifs with larger numbers of synaptic connections are only considered present in a network if they are unlikely to occur from random combinations of motifs with fewer connections. In previous work[45,46], only network-wide averages of these synaptic patterns were required, but here we derive motifs for each individual unit which sum to produce the full network-wide motif cumulants providing a neuron-by-neuron view of synaptic structure.

There are three main classes of synaptic patterns (**Fig. 4a**): 'Chain motifs' represent sequential synaptic connections. 'Convergent motifs' represent two neurons that project to a common downstream output neuron separated by one or many synapses. 'Divergent motifs' represent two neurons that share an upstream input unit separated by one or many synapses. The simplest motif is a 'first-order chain motif' which simply represents the average strength of synaptic inputs or outputs of a unit. To explain the responsiveness of excitatory output units we considered synaptic patterns shared with all four subpopulations in the network: 'output units' that project connections to the readout node, 'target responsive input units' that receive stimulus current on target trials, 'non-target responsive input units' that receive stimulus current on non-target trials, and 'inhibitory units' (**Fig. 4b**). For example, the first-

order chain motif from the inhibitory subpopulation are simply the average inhibitory synaptic inputs to a unit; the "2nd order divergent motif" between a unit and the inhibitory subpopulation would indicate that it receives synaptic inputs from the same units as inhibitory units (beyond what can be explained by chance).

Classically and non-classically responsive units demonstrated distinct patterns of connectivity to the four subpopulations of the network (output units, target responsive input units, non-target responsive input units, and inhibitory units). Examination of the monosynaptic motifs (direct synaptic inputs and outputs to the four subpopulations) revealed that non-classically responsive units received weaker inputs than classically responsive units overall, particularly from inhibitory units and target responsive units (**Fig. 4c,** monosynaptic inputs, change in median normalized synaptic weights for non-classically vs. classically responsive units = -0.20, $p < 10^{-5}$, Mann-Whitney U test two-sided; **Extended Data Fig. 7a, top**, all comparisons of NCR vs. CR $p < 10^{-5}$, Mann-Whitney U test two-sided Bonferroni-correction). Despite receiving weaker inputs**,** non-classically responsive units had stronger recurrent synaptic outputs than classically responsive units to all other excitatory subpopulations (**Fig. 4c,** monosynaptic outputs, change in median synaptic input weight for non-classically vs. classically responsive units = 0.13, $p < 10^{-5}$, Mann-Whitney U test two-sided; **Extended Data Fig. 7a, bottom**; all comparisons of NCR vs. CR except inhibitory outputs, $p < 10^{-4}$, inhibitory outputs $p > 0.7$, Mann-Whitney U test two-sided with Bonferroni-correction). Comparing disynaptic motifs, we found that only a limited number of disynaptic input motifs differed systematically from zero, indicating that most disynaptic connections can be explained by chance occurrences of monosynaptic motifs (**Extended Data Fig. 7b**). When differences were apparent, disynaptic input motifs were closer to zero for non-classically responsive units as compared to classically responsive units indicating that non-classically responsive units have less correlated synaptic inputs

overall (**Fig. 4c,** disynaptic inputs**,** change in median synaptic input weight for non-classically vs. classically responsive units = -0.36; **Extended Data Fig. 7b**, inputs $p < 0.005$ unless labeled n.s., Mann-Whitney U test two-sided Bonferroni-correction, outputs $p < 0.04$ unless labeled n.s., Levene's test with Bonferroni-correction).

Can we predict which neurons will become classically or non-classically responsive based on their local synaptic structure? To answer this, we formulated a statistical model to predict the firing rate modulation of individual units based on its individual synaptic motifs. This statistical model uses the prevalence of synaptic motifs for individual units to predict the stimulus and choice modulation of these units via a multilinear regression. These predictions are then combined to make a prediction for the overall firing rate modulation (**Fig. 4d**). This model successfully predicted the firing rate modulation of individual output neurons across various network types with radically different response profiles (pre-STDP, IE only, EE only, post-STDP) demonstrating that the response profile of individual units can be explained by their local patterns of synaptic weights independent of the plasticity mechanisms shaping those weights (**Extended Data Fig 7c**; 30-fold cross-validated stimulus mean-squared error = 0.29±0.15 spikes/s, r = 0.75±0.06, choice mean-squared error = 0.63±0.43 spikes/s, r = 0.87±0.02). Examining the output units as a whole, this statistical synaptic motif model captures the observed shift towards non-classically responsive activity post-STDP demonstrating that local synaptic plasticity can account for global changes in the response profile distribution (**Fig. 4e**, **Extended Data Fig. 7d**).

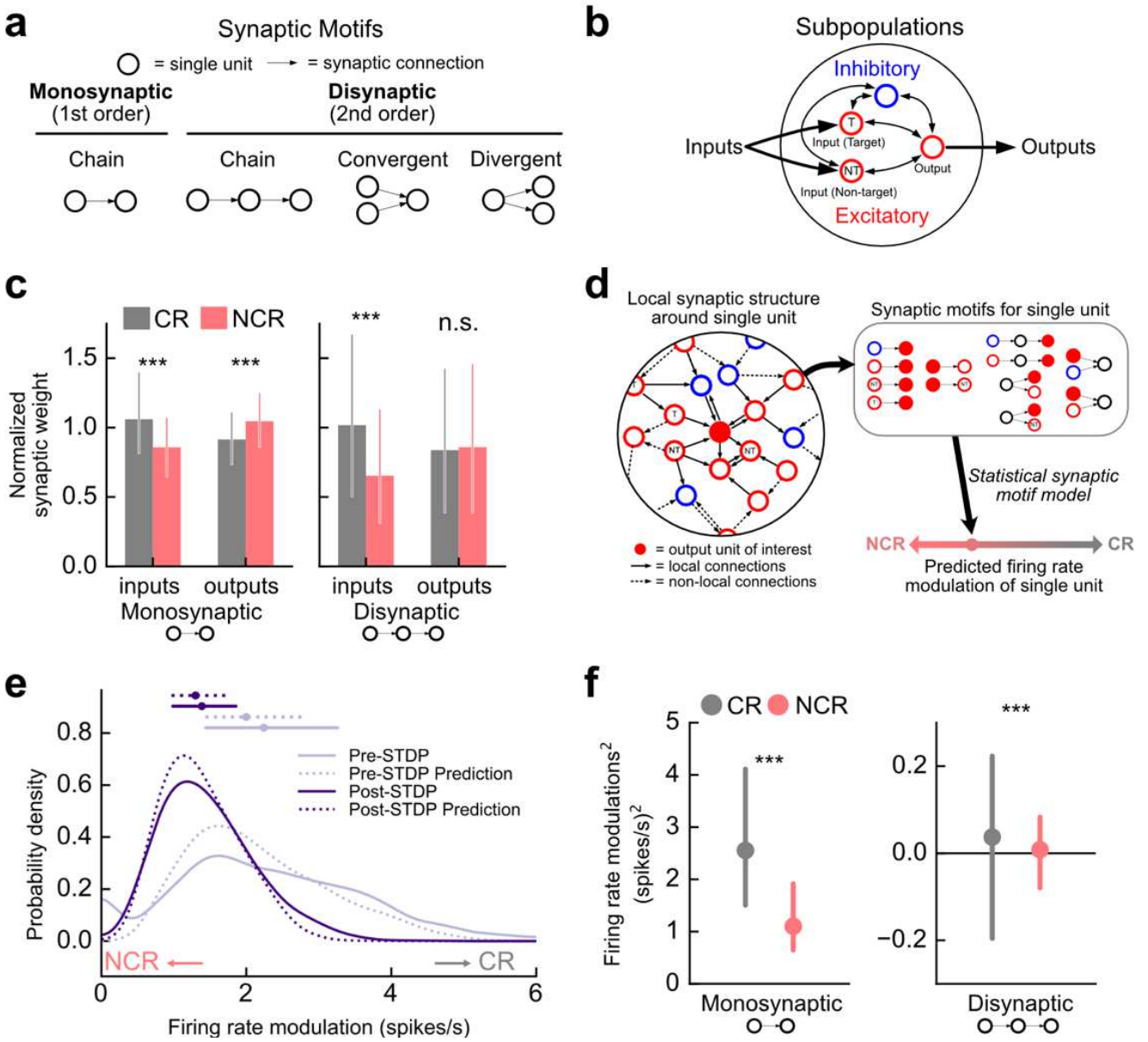

Figure 4. Specific local synaptic patterns predict response properties of diverse units. a, All possible monosynaptic (first-order) and disynaptic (second-order) motifs. **b,** Network schematic including 4 network subpopulations: target input units, non-target input units, output units, and inhibitory units. **c,** left, observed prevalence of all monosynaptic motifs between individual output units and all subpopulations for non-classically responsive units (NCR, red) and classically responsive units (CR, grey). Bars and lines represent medians and interquartile range, respectively. Right, same as left except for all disynaptic motifs between individual output units and all subpopulations. **d,** Schematic of method for predicting response profile from local synaptic structure around a unit. Local structure was decomposed into a set of synaptic motifs and these motifs were used to predict the modulation of the unit. **e,** Probability density of firing

rate modulation for individual output units for networks without STDP (pre-STDP, solid light purple) or all STDP rules (post-STDP, solid dark purple) along with predictions derived from statistical motif model (dotted lines). Small values of the firing rate modulation correspond to non-classical response profiles; high values correspond to classical response profiles. Summary circles and bars above distributions represent median and interquartile range, respectively. **f,** Contributions of individual monosynaptic motifs to single-unit firing rate modulation[2] for classically responsive units (CR, grey) and non-classically responsive units (NCR, red). Circles and bars represent median and interquartile range, respectively.

Which synaptic motifs are relevant for determining whether a neuron is classically or non-classically responsive? We dissected our statistical motif model to determine the contribution of each statistically significant monosynaptic and disynaptic motif to the firing rate modulation of classically responsive and non-classically responsive units. Monosynaptic motifs were primarily responsible for determining response profiles while disynaptic motifs played a smaller, non-systematic role (**Fig. 4f, Extended Data Fig. 7e,f**; all NCR vs. CR comparisons $p < 10^{-5}$ Mann-Whitney U test two-sided Bonferroni-correction). 6 out of a possible 8 monosynaptic motifs and 6 out of a possible 16 disynaptic motifs were found to be relevant for predicting the modulation of individual units indicating that a specific set of local synaptic motifs are relevant for identifying the response properties of network units (**Extended Data Fig. 7a,b,e-h**; $p < 0.01$, Mann-Whitney U test two-sided with Bonferroni-correction). Each of these identified motifs was relevant for the response properties of both classically responsive and non-classically responsive units. This indicated that both unit types were driven by the same types of connections. Moreover, these included connections to all subpopulations (output, target responsive input, non-target responsive input, and inhibitory) although higher-order, disynaptic connections to output and inhibitory units were of particular importance (**Extended Data Fig. 7e,f**; output, 3 of 4 possible disynaptic motifs, open red circles; inhibitory, 2 of 4 possible disynaptic motifs, open blue circles;). Decreases in monosynaptic inputs to non-classically responsive units and increases in their monosynaptic outputs (**Fig. 4c, Extended Data Fig. 7a**) all contributed to less modulated firing overall (**Fig. 4f, Extended Data Fig. 7e**). Moreover, the decrease in disynaptic correlations for non-classically responsive units (**Fig. 4c, Extended Data Fig. 7b**) resulted in a smaller disynaptic contribution to the response profiles of non-classically responsive units (**Fig. 4f, Extended Data Fig. 7f**).

Examining stimulus and choice related firing rate modulation separately revealed that there were distinctions in how each subpopulation affect specific task-related responses **(Extended Data Fig. 7g,h**). Monosynaptic connections to inhibitory and output units play a significant role in both stimulus and choice-related responses, however connections to input units only contribute significantly to stimulus responses. Furthermore, disynaptic connections play a relatively larger role in choice-related responses than stimulus responses indicating that higher-order structure may be more important in transforming stimulus-related into choice-related activity. Thus, local synaptic structure predicted the response profile of individual units with direct monosynaptic connections having the largest effect. The response properties of classically and non-classically responsive units were driven by differences in how they connect to all network subpopulations rather than one in isolation (**Fig. 4c,f, Extended Data Fig. 7a,b,e,f**). Specifically, non-classically responsive units resulted from weakened input from all subpopulations and smaller disynaptic correlations rather than increased inhibition or weakened inputs from input units. These differences were accompanied by an increase in the strength of projections from non-classically responsive units to the rest of the network.

Predicting single neuron response profiles recorded in vivo

Our model makes several predictions about the relationship between synaptic inputs and output spiking responses. Specifically, we hypothesized that neuronal response type (i.e., classically responsive or non-classically responsive) can be determined from average synaptic input strengths (**Fig. 4c-f**). We made in vivo whole-cell voltage-clamp recordings from neurons of the auditory cortex of mice during the go/no-go frequency recognition task (**Fig. 5a**; n=12 neurons from N=5 mice). We measured excitatory and inhibitory synaptic currents (E/IPSCs) during behavior, and for some cells (n=4 neurons from N=3 mice), we were able to record both the spiking activity in cell-attached mode prior to breaking into the cell to

record postsynaptic currents (**Fig. 5b,c, Extended Data Fig. 8a**). For those cells where only synaptic currents were recorded, we used a straightforward integrate-and-fire model to simulate their spiking activity based on the experimentally-measured currents over individual trials, including during inter-trial intervals (**Extended Data Fig. 8b**). The firing rate modulation of each neuron (and thus degree to which each neuron was classically responsive or non-classically responsive) was then calculated by comparing baseline spiking activity to activity during stimulus presentation. The parameters chosen for the integrate-and-fire simulation were based on those neurons where both synaptic input and spiking output data were available for the same neuron (**Fig. 5b,c**, right). This simulation accurately captured the firing rate modulation of cells where spiking data was available (**Fig. 5d**; Pearson's $r = 0.85$) and reproduced a distribution of responses similar to those directly measured from cell-attached recordings (**Fig. 5e**; Kolmogorov-Smirnov test, $p = 0.41$). These results confirm that our simulations based on recorded input currents accurately reproduce the modulation of neurons where spiking outputs were not available.

Our spiking RNN also predicts that non-classically responsive units have weaker average inputs overall. We also observed this in the whole-cell recordings in vivo; non-classically responsive neurons had significantly weaker inhibitory and excitatory inputs than classically responsive cells (**Fig. 5f**, NCR vs CR, $\Delta Exc = -72\%$, $p = 0.002$, $\Delta Inh = -76\%$, $p = 0.007$, Mann-Whitney two-sided U-test). We then used parameters derived from our RNN to predict the firing rate modulation of individual neurons in vivo using only these average inhibitory and excitatory conductances. These parameters were derived from a simplified version of the same statistical model previously used to predict modulation of units in the RNN using their synaptic motifs (**Fig. 4;** see Methods). The parameters were coefficients representing the relative contribution of average inhibitory and excitatory conductance to firing rate modulation during behavior and were combined with measured average synaptic conductances from cells in vivo to predict

their modulation (**Fig. 5g, left**). Comparing the modulation inferred directly from the trial-by-trial data with the one predicted by our model, we found the statistical model derived from the RNN predicts the firing modulation of individual neurons in vivo to a high degree of accuracy for neurons over a range of classically responsive and non-classically responsive firing rate modulations (**Fig. 5g,** right; mean-squared error = 2.2 spikes/s, Pearson's r = 0.94).

To test whether the RNN-derived predictions captured detailed aspects of the trial-by-trial dynamics, we compared these predictions to modified simulations which kept average conductance values fixed while rescaling trial-by-trial conductance dynamics relative to this baseline (peak conductance values were rescaled to be closer to the mean value by a factor of 2). This ensured that the RNN-based predictions would remain unchanged while simultaneously altering the trial-by-trial currents used for simulation. The modulations produced by these rescaled dynamics were systematically lower than the RNN-derived predictions indicating that the RNN-derived coefficients capture non-trivial features of the original trial-by-trial dynamics in vivo (**Extended Data Fig. 8c**, mean-squared error = 3.1 spikes/s, Pearson's r = 0.85). Next, we checked whether the relationship between synaptic inputs and output modulation derived from our model was statistically meaningful by randomly shuffling the modulation values for RNN units and deriving coefficients that attempt to predict these random values from the synaptic structure. Using these shuffled RNN coefficients on the experimental data significantly reduced the accuracy of the predicted modulations, demonstrating that the success of our RNN-derived predictions was not due to chance (**Extended Data Fig. 8d**, mean-squared error = 4.9 spikes/s, Pearson's r = 0.3). These findings indicate that our RNN model successfully recapitulates the connection between synaptic structure and spiking response properties observed in vivo.

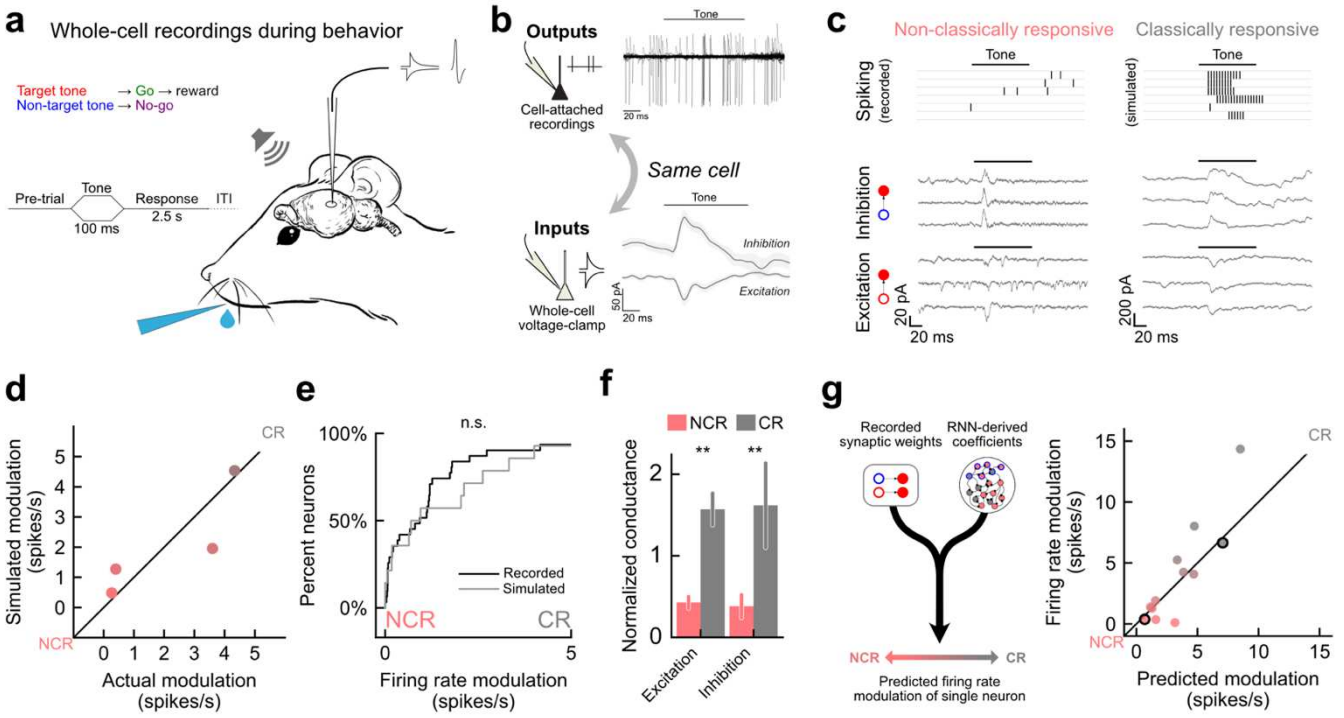

Figure 5. RNN-derived statistical motif model predicts *in vivo* response profiles. a, Schematic of behavioral task and whole-cell recording set-up during behavior. **b,** An example cell where cell-attached recording was used to first measure spiking outputs before breaking into the cell to record synaptic currents (E/IPSCs). **c,** Example recordings from a non-classically responsive neuron (left) and a classically responsive neuron (right). For non-classically responsive unit recorded spike times are shown; for classically responsive unit simulations are shown. **d,** Simulated firing rate modulation versus actual modulation for 4 neurons where spiking outputs and synaptic inputs were both recorded (Pearson's r = 0.85). **e,** Comparison of cumulative distribution function from cell-attached recordings (black) and leaky-integrate-and-fire simulation (grey). No significant difference was observed (Kolmogorov-Smirnov test, p = 0.41) **f,** Comparison of average synaptic inputs to non-classically responsive neurons (red) versus classically responsive neurons. Bars indicate means and S.E.M. (NCR vs CR, ΔExc = -72%, p = 0.002, ΔInh = - 76%, p = 0.007, Mann-Whitney two-sided U-test). **g,** recorded/simulated modulation was compared to predictions based on coefficients from the RNN-derived motif statistical model, neurons from **c** circled in black. (n=12, mean-squared error = 2.2 spikes/s, Pearson's r = 0.94).

Discussion

Our experimentally-motivated spiking RNN model revealed that the diversity of neuronal response profiles is constrained by local synaptic structures and shaped by synaptic plasticity mechanisms. This model sits at the nexus between two recent trends in neural network modeling: First, recent work has successfully extended general-purpose learning algorithms (e.g., FORCE) designed for rate-based networks to networks with spiking units[35,47–49]. Second, there has been renewed interest in using RNNs to understand the role of biologically-motivated synaptic plasticity rules in the formation of neuronal ensembles important for learning, memory storage, and task performance[32,33,50]. We showed how these two methods can cooperate to shape the synaptic organization underlying heterogeneous neuronal responses for performing a classification task. This produced a 'synaptic signature' in terms of the monosynaptic connections received by any given neuron, related to how that cell was wired into the network, which learning rules helped shape those connections, and the function of that cell for task performance.

In principle and in practice, any neuron – or even every neuron – can contribute in some way to overall network dynamics and behavioral performance, regardless as to the degree of overt firing rate changes (i.e., classical responsiveness). Both classically and non-classically responsive units directly affected network task performance, but via distinct mechanisms. Diverse neuronal responses throughout a population are likely essential for robustly supporting the dynamics required for sensorimotor transformations and adaptive behaviors. This diversity might also enable evidence accumulation, decision making, working memory[14], and learning in spiking RNNs[51], enabling neural processing and flexible computations over a range of spatiotemporal scales throughout the brain.

Methods

Electrophysiological recordings during behavior

All animal procedures were performed in accordance with National Institutes of Health standards and were conducted under a protocol approved by the New York University School of Medicine Institutional Animal Care and Use Committee. For single-unit recordings in freely moving rats, 5 adult male and 6 adult female Sprague-Dawley rats were trained to perform a go/no-go frequency recognition task as previously described[15,52,53]. Animals were trained to respond via nosepoke to a target tone (4 kHz) for food reward and to withhold their response to nontarget tones (0.5,1,2,8,16,32 kHz). The response period was 2.5 sec and false alarms resulted in a 7 sec timeout. All pure tones played were 100 ms duration, 3 ms cosine on/off ramps, at 70 dB sound pressure level (SPL). Behavioral events (stimulus delivery, food delivery, and nosepoke detection) were monitored and controlled with a custom-programmed microcontroller (Med Associates). After animals reached behavioral criteria (percent correct: ≥70%, d': ≥1.5), rats were anesthetized with ketamine (40 mg/kg) and dexmedetomidine (0.125 mg/kg) and implanted with tetrode microdrive arrays (Versadrive-8 Neuralynx) in right auditory cortex as previously described[15]. Tetrodes were advanced the day prior to recordings, signals were filtered between 250 Hz and 5 kHz, digitized at 30 kHz and spikes were sorted using Offline Sorter (Plexon Inc)[15,52]. The animals and units shown were previously described[15].

For cell-attached recordings in head-fixed mice, 4 adult male and 6 adult female C57Bl/6 mice were anesthetized with isoflurane (3% during induction, 2% during surgery), and a custom-designed stainless steel headpost was affixed to the skull with dental cement (Metabond). Following 7+ days of recovery and 7+ days of subsequent water-restriction, mice were trained on a go/no-go frequency-recognition task[6]. Animals were trained to respond to the target tone (11.2 kHz) by licking for water reward and to withhold

responses to the non-target tone (5.6 kHz). Acoustic stimuli were 100 ms duration, 3 ms cosine on/off ramps, at 70dB SPL. Animals had 2.5 seconds to respond and false alarms resulted in a 7 sec timeout. Behavioral events (stimulus delivery, water delivery, and lick detection) were monitored and controlled by custom-written programs in MATLAB that interfaced with an RZ6 processor (Tucker-Davis Technologies). After animals reached behavioral criteria (percent correct: ≥70%, and d': ≥1.5), mice were anesthetized with isoflurane and a 3 mm diameter glass cranial window with a small 200 μm hole to allow for pipette access was implanted over auditory cortex (1.75 mm anterior to the lambdoid suture). After recovery, *in vivo* cell-attached or whole-cell recordings were obtained from neurons located 300-900 μm below the pial surface during behavior as previously described[6]. Recordings were made in a sound-attenuation chamber (Eckel) using a Multiclamp 700B amplifier (Molecular Devices). For voltage-clamp experiments, whole-cell pipettes (5-7 MΩ) contained (in mM): 130 Cs-methanesulfonate, 4 TEA-Cl, 4 MgATP, 10 phosphocreatine, 10 HEPES, 1 QX-314, 0.5 EGTA, 2 CsCl, pH 7.2, R_i = 283 ± 152 MΩ (s.d.). Data were filtered at 2 kHz, digitized at 20 kHz, and analyzed with Clampfit 10 (Molecular Devices). Cells were held at -70mV to measure EPSCs and above 0 mV for IPSCs.

Characterization of response profiles using a continuous measure

To comprehensively characterize spiking responses, we used a continuous measure of responsiveness which generalizes the binary classification (classically vs. non-classically responsive) we used previously[15]. Our continuous measure of responsiveness quantified the degree to which a cell exhibited firing rate changes during both the stimulus and choice periods. For the experimental data, we calculated the trial-averaged change in firing rate for each neuron during stimulus presentation relative to intertrial baseline, R_{st}, using a 150 ms window from stimulus onset to 50 ms post-stimulus to capture offset responses. The trial-averaged change in firing rate prior to behavioral choice, R_{ch}, was calculated using a

window spanning the 500 ms prior to behavioral response on 'go' trials and 500 ms prior to the average behavioral response on 'no-go' trials. For spiking RNN units, stimulus modulation was calculated using the 100 ms stimulus period and choice modulation was calculated using the 100 ms choice period.

Our overall measure of firing rate modulation, R, combined these two terms in quadrature so that both stimulus- and choice-related firing rate changes must be small for a unit to be characterized as having a low firing rate modulation (i.e., have a more non-classically responsive response profile),

$$R^2 = R_{\mathrm{st}}^2 + R_{\mathrm{ch}}^2. \quad (2)$$

This measure captures the detailed firing rate modulation of individual units during both the stimulus and response periods. It provides information about the degree to which a unit is classically responsive such that values close to 0 are only possible when a unit is non-classically responsive (e.g., a unit with firing rate modulation of 0.1 spikes/s would be classified as a non-classically responsive unit whereas we would consider 5 spikes/s as highly classically responsive).

Discrete characterization of classically and non-classically responsive units

Statistical identification of classically and non-classically responsive units followed our previous methods[15]. We used two positive statistical tests for non-classical responses to establish lack of responses during either the stimulus and/or response periods. The test compared the number of spikes during each of these windows to inter-trial baseline. Given that spike counts are discrete, bounded, and non-normal, we used subsampled bootstrapping to evaluate whether the mean change in spikes during tone presentation or the response period was sufficiently close to zero (in our case, 0.2 spikes). We subsampled 90% of the spike count changes from baseline, calculated the mean of these values, and repeated this process 5000 times to construct a distribution of means. If 95% of the subsampled mean values were between -0.2 and

0.2, we considered the cell sensory non-classically responsive ($p<0.05$). This is a conservative, rigorous method for identifying a cell or unit as being 'non-classically responsive'[15].

Spiking recurrent neural network model

To study the origin and functional contributions of diverse neural response profiles, we simulated a spiking neural network of 1,000 sparsely connected (5% connection probability) leaky integrate-and-fire units (800 excitatory, 200 inhibitory) with current-based synaptic input. All parameters listed in **Extended Data Table 1**. The temporal evolution of the membrane voltage V_i of unit i is

$$\tau_m \frac{dV_i}{dt} = V_i - V_r + I_{Ei}(t) - I_{Ii}(t) + I_0 \quad (3)$$

where τ_m = 20 ms is the membrane time constant, V_r = -65 mV is the resting membrane potential, $I_{Ei}(t)$ and $I_{Ei}(t)$ are the recurrent EPSCs and IPSCs to unit i respectively, and I_0 is the leak current. Upon reaching a threshold value of V_{th} = -55 mV the unit emits an action potential and its membrane voltage is reset to V_r .

EPSCs and IPSCs decay exponentially with time constants of τ_E = 20 ms and τ_I = 20 ms respectively such that if neuron j synapses onto neuron i with synaptic weights W_{ij} and fires action potentials at times $\{t_{jk}\}$ the excitatory and inhibitory currents are:

$$I_{Ei}(t) = \sum_{k,j \in E} W_{ij}(t_{jk})\, \Theta(t - t_{jk})\, e^{-(t-t_{jk})/\tau_E}$$
$$I_{Ii}(t) = \sum_{k,j \in I} W_{ij}(t_{jk})\, \Theta(t - t_{jk})\, e^{-(t-t_{jk})/\tau_I} \quad (4)$$

Where $\boldsymbol{\Theta}$ is the Heaviside step function. Non-zero values of the weight matrix W_{ij} were initialized to a uniform distribution between 0 and a maximum value. The initial mean value was set differently for each connection type in our network (excitatory-to-excitatory, inhibitory-to-excitatory, excitatory-to-inhibitory, inhibitory-to-inhibitory) to ensure initial chaotic dynamics amenable to FORCE training (see

below ‘Task and FORCE training’). Initial output weights scaled by the square root of the average number of incoming connections from each subpopulation (excitatory or inhibitory):

$$W_{E/I} = \frac{W_{0\,E/I}}{\sqrt{p_{con} N_{E/I}}}. \quad (5)$$

Excitatory-to-inhibitory and inhibitory-to-inhibitory weights remained fixed throughout simulation. In contrast, excitatory-to-excitatory and inhibitory-to-excitatory weights were modified by a form of spike timing plasticity (see below ‘Synaptic plasticity mechanisms’).

Task and FORCE training

The network was trained on a go/no-go stimulus classification task similar to that used experimentally for rats and mice[15,52,53]. During a 100 ms stimulus period, the network was stimulated with one of seven possible inputs (corresponding to frequencies of 0.5, 1, 2, 4, 8, 16, 32 kHz), and trained to produce an output during the subsequent 100 ms response period if the input was the 4 kHz target tone, while remaining at baseline for all other non-target tones. In between each 200 ms trial were inter-trial intervals randomly chosen from a uniform distribution between 100-400 ms.

A subset of excitatory units were designated as input units (N_{in} = 200) that received additional current during the stimulus period. Our stimuli were represented using a place code. For each stimulus s a fixed subset of units G_s received an additional fixed current I_{in} while all other input units received no additional current:

$$I_{\mathrm{in}\,j} = \begin{cases} I_{\mathrm{in}} & \text{if } j \in G_s \\ 0 & \text{if } j \notin G_s \end{cases}. \quad (6)$$

The input current magnitude, I_{in}, was minimized while ensuring that the network could still perform the task. The remaining excitatory units were output units (N_{out} = 600) which projected an additional set of

weights, w, to a readout node whose activity, $z_{out}(t)$, represents the response of the network. The activity of the readout node is a weighted sum of the activity of all output units smoothed by an exponential kernel with time constant τ_{out} = 100 ms. If output unit i fires action potentials at times $\{t_{ik}\}$, its smoothed output is:

$$s_i(t) = \sum_k \frac{1}{\tau_{\text{out}}} \Theta(t - t_{ik})\, e^{-(t-t_{ji})/\tau_{\text{out}}} \tag{7}$$

and the readout node activity is calculated as:

$$z_{\text{out}}(t) = \sum_{i \in E_{\text{out}}} w_i(t)\, s_i(t) \tag{8}$$

.

The network was trained to produce a burst in readout node activity when the target stimulus was presented ($s = T$) and remain at baseline for all other stimuli ($s \neq T$) using a version of FORCE training adapted to spiking recurrent neural networks[34,35]. Specifically, $z_{out}(t)$ was trained to approximate:

$$f_{\text{out}}(t) = \begin{cases} \Theta(t-100)\, \sin\left(\frac{(t-100)\pi}{100}\right) & s = T \\ 0 & s \neq T \end{cases} \tag{9}$$

.

FORCE training requires that network output units receive feedback from the readout node in the form of an additional current. If $I_{FB\,i}(t)$ is the feedback current to readout node i then

$$I_{\text{FB}\,i}(t) = Q\, \eta_i\, z_{\text{out}}(t) \tag{10}$$

Where Q is the network-wide feedback strength, η_i is the fixed feedback weight onto output unit i chosen from a uniform distribution between -1 and 1.

During FORCE training, the network iteratively modified the output weights w_i to reduce the error between the network output and the desired output function: $e(t) = f_{out}(t) - z_{out}(t)$. Updates to the output weights were made at random times with an average interval of T_{FORCE} = 4 ms. Using bold font to denote vectors and matrices, the learning rule for a given updated interval of Δt is:

$$\boldsymbol{w}(t) = \boldsymbol{w}(t - \Delta t) - \frac{e(t)\boldsymbol{P}(t)\boldsymbol{s}(t)}{1 - \boldsymbol{s}^T(t)\boldsymbol{P}(t)\boldsymbol{s}(t)} \tag{11}$$

where *P(t)* is the network estimate for the inverse of the correlation matrix and is updated according to:

$$\boldsymbol{P}(t) = \boldsymbol{P}(t - \Delta t) - \frac{\boldsymbol{P}(t - \Delta t)\boldsymbol{s}(t)\boldsymbol{s}(t)^T\boldsymbol{P}(t - \Delta t)}{1 + \boldsymbol{s}(t)^T\boldsymbol{P}(t - \Delta t)\boldsymbol{s}(t)} \tag{12}$$

Initial weights, *w(0)*, were chosen from normally distributed gaussian values with a standard deviation of $k_0 = 0.1 / N_{out}$ and P(0) = *I* where *I* is the identity matrix and λ is a model parameter which acts as a regularizer such that:

$$\boldsymbol{P}(t)^{-1} = \int_0^t \boldsymbol{s}(t)\boldsymbol{s}(t)^T \mathrm{dt} + \lambda \boldsymbol{I} \tag{13}$$

In our network, $\lambda = 1$. Once training was complete output weight *w* remained fixed.

After output weights stabilized and networks reached asymptotic performance (i.e., after 2000 to 5000 trials), all weights were fixed and additional trials were run to evaluate network performance. Network performance was calculated by examining the integrated output activity during the response period (100 to 200 ms):

$$R = \int_{100}^{200} z_{\mathrm{out}}(t)\ \mathrm{d}t \tag{14}$$

An optimal threshold was calculated to distinguish the integrated output on target versus non-target trials by regressing the trial type (target vs. non-target) against the integrated output activity, *R*, using logistic regression. Integrated outputs above/below this threshold were considered go/no-go responses. All d' were calculated using these response designations.

Synaptic plasticity mechanisms

FORCE training modified the output weight vector *w* while the recurrent weight matrix was modified by homo- and heterosynaptic forms of STDP. All recurrent inputs to excitatory cells were plastic and we used

distinct homosynaptic mechanisms for excitatory-to-excitatory synapses and inhibitory-to-excitatory synapses. Inputs to inhibitory units remained fixed throughout training and testing.

Homosynaptic excitatory-to-excitatory plasticity

In our model, homosynaptic excitatory-to-excitatory plasticity follows a standard Hebbian model[36,54] in that when a presynaptic cell fired before a post synaptic cell, long-term potentiation (LTP) occurred and when a postsynaptic cell fired before a presynaptic cell, long-term depression (LTD) occurred. The rule can be expressed as

$$\begin{aligned} \Delta W_{ij}(t) &= A\, W_{ij}(t)\, z_{+\,j}(t) && \text{when } \mathbf{post}\text{synaptic cell } i \text{ fires} \\ \Delta W_{ij}(t) &= -B\, W_{ij}(t)\, z_{-\,i}(t) && \text{when } \mathbf{pre}\text{synaptic cell } j \text{ fires} \end{aligned} \tag{15}$$

Where $z_{+j}(t)$ and $z_{-j}(t)$ are each increased by 1 each time neuron j fires an action potential and exponentially decay with a time constant of $\tau_+ = \tau_- = 20$ ms and $A = 0.001$ and $B = 0.00105$ represent the strength of LTP and LTD respectively.

Homosynaptic inhibitory-to-excitatory plasticity

Inhibitory-to-excitatory homosynaptic plasticity was based on experimental[38] and theoretical[39] work demonstrating that synapses between inhibitory and excitatory auditory cortical cells undergo LTP when spike timing is synchronous regardless of order and undergo LTD when firing is asynchronous,

$$\begin{aligned} \Delta W_{ij}(t) &= \eta_I\, W_{ij}(t)\, (z_{I\,j}(t) - \alpha) && \text{when } \mathbf{post}\text{synaptic cell } i \text{ fires} \\ \Delta W_{ij}(t) &= \eta_I\, W_{ij}(t)\, z_{I\,i}(t) && \text{when } \mathbf{pre}\text{synaptic cell } j \text{ fires}, \end{aligned} \tag{16}$$

where $\eta_I = 0.001$ is the overall strength of inhibitory-to-excitatory plasticity, α represents the strength of LTD (see "Training Protocol" for more information). $z_{I\,j}(t)$ increases by 1 each time neuron j fires an action potential and exponentially decays with a time constant of $\tau_I = 5$ ms. This time constant was set to match the crossover time scale of LTP to LTD observed experimentally[38].

Heterosynaptic inhibitory-to-excitatory and excitatory-to-excitatory plasticity

Given the instability of pairwise Hebbian excitatory plasticity[32,55], we included two forms of heterosynaptic plasticity on excitatory-to-excitatory and inhibitory-to-excitatory synapses based on previous spiking RNN studies[33] which would (1) systematically weaken all presynaptic weights to prevent any one presynaptic connection from dominating (heterosynaptic balancing)

$$\Delta W_{ij}(t) = -\beta\, W_{ij}(t)\,(z_{-\,j}(t))^3 \qquad \text{when } \mathbf{post}\text{synaptic cell } i \text{ fires} \tag{17}$$

and (2) systematically strengthen postsynaptic weights to prevent a weakened synapse from dropping out entirely (heterosynaptic balancing)

$$\Delta W_{ij}(t) = \delta \qquad \text{when } \mathbf{pre}\text{synaptic cell } j \text{ fires}. \tag{18}$$

Training protocol

In general, networks were trained until output weights, recurrent weights, and network behavior stabilized (2000 – 5000 trials). In addition to the dynamic mechanisms listed above the network-wide bias current, I_0, was tuned so that inhibitory firing was on average $r_I \approx 20$ spikes/s across the population. This was accomplished via a learning rule which adjusted the bias current after each trial by

$$\Delta I_0 = \eta_r\,(R_I - r_I) \tag{19}$$

Where $\eta_r = 0.005$ is the bias current learning rate and R_I is the average firing rate of all inhibitory cells.

Given these three plasticity mechanisms (STDP, FORCE, and bias current adjustment) training proceeded in a staged manner to ensure learning for each mechanism occurred in a regime where major initial changes induced by other mechanisms had subsided and subsequent plasticity could be regarded as quasi-stable. (1) Initially only bias current adjustment was active in order to set inhibitory rates to their desired

firing rate prior to activating activity dependent plasticity mechanisms (STDP). (2) After 40 trials, all STDP mechanisms were activated, and bias currents were permitted to keep adjusting to maintain inhibitory rates near their target value. During this period, inhibitory plasticity adjusted excitatory firing rates towards a target value set, r_E = 10 spikes/s. This value is determined via the strength of inhibitory-to-excitatory LTD (Eqn. 15) by the equation[39]

$$r_E = \frac{\alpha}{2\tau_I}. \tag{20}$$

(3) After an additional 60 trials (trial 100), FORCE training was activated once inhibitory and excitatory rates had reached their target values. STDP and bias current adjustment was maintained. (4) After an additional 100 trials (trial 200), bias current adjustment ceased and STDP and FORCE training remained active simultaneously for the remainder of the training session. The observed boost in performance occurred when STDP and FORCE were simultaneously active for at least ~1000 trials (**Extended Data Fig. 1d**) indicating that the two mechanisms can interact productively.

Simulation details and code

All simulations were carried out via an event-based simulator which integrated all dynamical variables between spiking events. All simulations were all conducted in Julia. See Brette et al. for a description of the method and comparison with other methods[56] and Engelken for implementation details[57].

Single-trial, ISI-based Bayesian decoding

We applied a single-trial Bayesian ISI-based trial-by-trial decoding algorithm previously described[15]. In brief, using a training set composed of 90% of the data recorded trials the probability density function for observing an ISI on target/nontarget trials (or go/no-go trials) was inferred (p(target), p(non-target), p(go), and p(no-go)) via kernel density estimation with the bandwidth set by cross-validated maximum likelihood estimation. These probability density functions were used to infer the probability of a stimulus and choice

from the observed ISIs, {ISI} , on a new trial taken from the remaining 10% of test trials via Bayes rule (assuming statistical independence between the ISIs observed)

$$\begin{aligned} \mathrm{p}(\mathrm{stimulus}|\{\mathrm{ISI}\}) &= \frac{\prod_i \mathrm{p}(\mathrm{ISI}_i|\mathrm{stimulus})\,\mathrm{p}(\mathrm{stimulus})}{\sum_{\mathrm{stimulus}} \prod_i \mathrm{p}(\mathrm{ISI}_i|\mathrm{stimulus})\,\mathrm{p}(\mathrm{stimulus})}, \\ \mathrm{p}(\mathrm{choice}|\{\mathrm{ISI}\}) &= \frac{\prod_i \mathrm{p}(\mathrm{ISI}_i|\mathrm{choice})\,\mathrm{p}(\mathrm{choice})}{\sum_{\mathrm{choice}} \prod_i \mathrm{p}(\mathrm{ISI}_i|\mathrm{choice})\,\mathrm{p}(\mathrm{choice})}. \end{aligned} \tag{21}$$

Small-world path length analysis

To calculate the shortest path length shown in **Extended Data Fig. 3c**, we reduced our synaptic connection matrix W by setting a weight threshold, culling connections whose weights fall below said threshold and using the remaining weights to define the adjacency matrix. To calculate the mean shortest path lengths of the network, we performed a breadth-first search from each neuron along its outgoing connections modified to also calculate the shortest path from a neuron to itself (either a self-edge or a cycle in the network). The mean of all distances across all starting neurons is taken as the mean shortest path length for a given culling thresholds.

We compared these results against the same calculation performed on Watts-Strogatz small-world graphs which contain a mixture of regular local structure and random global structure. Previous work has suggested that cortical neural networks display connection statistics (such as mean path length) more consistent with a small-world network rather than one with purely random global connections[42,44]. Watts-Strogatz small-world networks vary parametrically from entirely local structure to entirely global structure via a parameter β spanning from 0 (completely regular ring-lattice) to 1 (completely random connections between vertices). Note that this parameter is unrelated to the heterosynaptic balancing parameter β. To compare the culled adjacency matrix from our model to a Watt-Strogatz small-world network, we match the number of edges in the Watts-Strogatz graphs to the number of connections in the culled networks and

generate the small-world network by first generating a bidirectional regular ring lattice with the same number of unidirectional edges as the adjacency matrix derived from our network and assign each unidirectional edge to a random endpoint with chance β.

Inactivation experiments

Because recurrent and output connections play qualitatively different roles in network dynamics we used separate procedures to inactivate each type of connection. Full inactivation of a unit employed output and recurrent inactivation procedures simultaneously. To inactivate the output contributions of a neuron, its output weight was held at zero so that its activity made no contribution to the activity of the readout node. To inactivate a neuron's recurrent contributions, the neuron was not silenced but rather its effect on postsynaptic neurons was replaced by a Poisson process that fired with a fixed rate equivalent to that of the neuron that was being replaced. Using this approach, we could be certain that the deficits observed were not caused by an overall decrease in network synaptic current (the original unit and Poisson unit fired at the same average rate) but could instead be attributed to the removal of all spike timing information present in the unit's firing pattern.

Inactivation targeting classical/non-classical subpopulations proceeded by characterizing the firing rate modulation (see "Characterization of response profiles using a continuous measure") and inactivating in order from most to least classically/non-classically responsive unit. For example, inactivating 10% of output units during the experiment targeting the non-classically responsive subpopulation would correspond to inactivating the top 10^{th} percentile of non-classically responsive units. Variability during the choice period was evaluated by calculating the root mean squared error between the readout node activity and correct output function on each trial.

Synaptic motif cumulant calculations

Synaptic motif cumulants represent the extent to which a pattern of synaptic connections (e.g., a disynaptic chain) between two subpopulations (in our case, input target units, input nontarget units, output units, or inhibitory units) are present relative to chance combinations of motifs with fewer synaptic connections[45,46]. Recent work has used calculated average motif cumulants across the network, but here we extend this work by examining motif cumulants for each network unit on an individual basis to examine how these synaptic patterns relate to the response profiles of individual units.

A synaptic motif is defined as the average product of synaptic weights present in a specified pattern of connections. For example, the n-chain motif, μ_n^{ch},is the average product of synaptic weights in a n-synapse chain of unidirectional connections from one unit to another and is calculated using the expression

$$\mu_n = \frac{\sum_{i,j} W_{ij}^n}{N^{n+1}} = \frac{\langle \boldsymbol{W^n} \rangle}{N^{n-1}}. \tag{22}$$

Where W is the synaptic weight matrix, N is the number of network units, and $\langle \cdot \rangle$ denotes the average over all matrix elements. The two other motifs shown to be significant for network dynamics are the (m,n)-diverging and (m,n)-converging motifs, $\mu_{m,n}^{di}$and $\mu_{m,n}^{co}$, which describe when two chains (of length m and n) diverge from the same unit or converge onto the same unit. For 2nd order motifs (disynaptic) these are the only other two possible motifs beyond chain motifs and are calculated

$$\begin{aligned} \mu_{m,n}^{\mathrm{di}} &= \frac{\langle \boldsymbol{W}^m \boldsymbol{W}^{T\,n} \rangle}{N^{m+n-1}}, \\ \mu_{m,n}^{\mathrm{co}} &= \frac{\langle \boldsymbol{W}^{T\,m} \boldsymbol{W}^n \rangle}{N^{m+n-1}}. \end{aligned} \tag{23}$$

To decompose these motifs into motif cumulants we introduce

$$\begin{aligned}\boldsymbol{u} &= (1, 1.., 1)^T/\sqrt{N},\\ \tilde{\boldsymbol{u}} &= (1, 1.., 1)^T/N,\\ \boldsymbol{H} &= \boldsymbol{u}\boldsymbol{u}^T,\\ \boldsymbol{\Theta} &= \boldsymbol{I} - \boldsymbol{H},\\ \boldsymbol{W}_\theta^n &= (\boldsymbol{W}\Theta)^{n-1}\boldsymbol{W}.\end{aligned} \tag{24}$$

and define the n-chain, (m,n)-diverging, and (m,n)-converging motif cumulants as

$$\begin{aligned}\kappa_n &= \frac{\tilde{\boldsymbol{u}}^{\boldsymbol{T}}\boldsymbol{W}_\theta^n\tilde{\boldsymbol{u}}}{N^{n-1}}\\ \kappa_{m,n}^{\mathrm{di}} &= \frac{\tilde{\boldsymbol{u}}^{\boldsymbol{T}}\boldsymbol{W}_\theta^m\Theta\boldsymbol{W}_\theta^{n\,T}\tilde{\boldsymbol{u}}}{N^{m+n-1}},\\ \kappa_{m,n}^{\mathrm{co}} &= \frac{\tilde{\boldsymbol{u}}^{\boldsymbol{T}}\boldsymbol{W}_\theta^{m\,T}\Theta\boldsymbol{W}_\theta^n\tilde{\boldsymbol{u}}}{N^{m+n-1}}.\end{aligned} \tag{25}$$

The motifs can then be decomposed in terms of the motif cumulants as

$$\begin{aligned}\mu_n &= \sum_{\{n_1,\ldots,n_t\}\in\mathcal{C}(n)}\left(\prod_{i=1}^{i}\kappa_{n_i}\right)+\kappa_n,\\ \mu_{m,n}^{\mathrm{di}} &= \sum_{\substack{\{n_1,\ldots,n_t\}\in\mathcal{C}(n)\\ \{m_1,\ldots,m_s\}\in\mathcal{C}(m)}}\left(\prod_{i=2}^{i}\kappa_{m_i}\right)\left(\kappa_{m_1,n_1}^{\mathrm{di}}+\kappa_{m_1}\kappa_{n_1}\right)\left(\prod_{i=2}^{s}\kappa_{n_j}\right)+\kappa_{m,n}^{\mathrm{di}},\\ \mu_{m,n}^{\mathrm{co}} &= \sum_{\substack{\{n_1,\ldots,m_t\}\in\mathcal{C}(n)\\ \{m_1,\ldots,m_s\}\in\mathcal{C}(m)}}\left(\prod_{i=2}^{i}\kappa_{m_i}\right)\left(\kappa_{m_1,n_1}^{\mathrm{co}}+\kappa_{m_1}\kappa_{n_1}\right)\left(\prod_{i=2}^{s}\kappa_{n_j}\right)+\kappa_{m,n}^{\mathrm{co}}\end{aligned} \tag{26}$$

where $C(n)$ represents all possible sets $\{n_1, \ldots, n_t\}$ of non-zero integers such that $\sum_{i=1}^{t} n_i = n$ excluding the set containing only one element $\{n\}$. The sum represents all possible ways of constructing the motif from lower order cumulants so we can interpret the final term on the right hand side as the contribution that cannot be attributed to lower order terms.

The equations above can be generalized if the network is composed of N units divided into k subpopulations. We specify each subpopulation P_q of size N_q (in our case, target input units, non-target-input units, output units, and inhibitory units) using u_q, a column vector of length 1 with equal entries for the components corresponding to members of the subpopulation and 0 for all other components. In other words, for subpopulation i and unit j the components of vector u_q are

$$u_{qj} = \begin{cases} 1/\sqrt{N_q} & \text{for } j \in P_q \\ 0 & \text{for } j \notin P_q \end{cases}. \tag{27}$$

We then specify the $N \times k$ matrix, U, as the concatenation of these subpopulation vectors,

$$\boldsymbol{U} = [\boldsymbol{u}_1 \dots \boldsymbol{u}_k], \tag{28}$$

which in component notation is simply $U_{jq} = u_{qj}$. Similarly, we define $\widetilde{U}$ via

$$\tilde{u}_{qj} = \begin{cases} 1/N_q & \text{for } j \in P_q \\ 0 & \text{for } j \notin P_q \end{cases}. \tag{29}$$

We now define a matrix of subpopulation motifs corresponding to the average product of synaptic weights present in specified pattern of connections where the first and last units of the motif are in subpopulation r and q respectively,

$$\begin{aligned} \mu_{n\,qr} &= \frac{\langle \boldsymbol{W}^{\boldsymbol{n}} \rangle_{qr}}{N^{n-1}}, \\ \mu^{\mathrm{di}}_{m,n\,qr} &= \frac{\langle \boldsymbol{W}^{m} \boldsymbol{W}^{T\,n} \rangle_{qr}}{N^{m+n-1}}, \\ \mu^{\mathrm{co}}_{m,n\,qr} &= \frac{\langle \boldsymbol{W}^{T\,m} \boldsymbol{W}^{n} \rangle_{qr}}{N^{m+n-1}}. \end{aligned} \tag{30}$$

Where $\langle \cdot \rangle_{qr}$ denotes the average over initial subpopulation r and final subpopulation q. In matrix form we write the motifs as $k \times k$ matrices

$$\begin{aligned}\boldsymbol{\mu}_n &= \{\mu_{n\,qr}\},\\ \boldsymbol{\mu}_n^{\mathrm{di}} &= \{\mu_{n\,qr}^{\mathrm{di}}\},\\ \boldsymbol{\mu}_n^{\mathrm{co}} &= \{\mu_{n\,qr}^{\mathrm{co}}\}.\end{aligned} \tag{31}$$

As expected, these subpopulation motifs can be used to calculate the total motifs using a weighted average. For example, the n-chain motif is calculated from the subpopulation motifs

$$\mu_n = \sum_{q,r} \frac{N_q N_r}{N^2} \mu_{n\,qr} \tag{32}$$

.

As in the single population case, these subpopulation motifs can be broken down into a series of subpopulation motif cumulants. To calculate the subpopulation motif cumulants we follow the same process as the single population case. Defining

$$\begin{aligned}\boldsymbol{H} &= \boldsymbol{U}\boldsymbol{U}^T,\\ \boldsymbol{\Theta} &= \boldsymbol{I}_N - \boldsymbol{H},\\ \boldsymbol{W}_\theta^n &= (\boldsymbol{W}\Theta)^{n-1}\boldsymbol{W}.\end{aligned} \tag{33}$$

The $k \times k$ matrix of average n-chain cumulants between each subpopulation is then calculated as

$$\boldsymbol{\kappa}_n^{\mathrm{ch}} = \frac{1}{N^{n-1}} \tilde{\boldsymbol{U}}^T \boldsymbol{W}_\theta^n \tilde{\boldsymbol{U}} \tag{34}$$

and (m,n)-divergent and (m,n)-convergent motif cumulants are calculated

$$\begin{aligned}\boldsymbol{\kappa}_{m,n}^{\mathrm{di}} &= \frac{1}{N^{m+n-1}} \tilde{\boldsymbol{U}}^T \boldsymbol{W}_\theta^m \Theta \boldsymbol{W}_\theta^{n\,T} \tilde{\boldsymbol{U}},\\ \boldsymbol{\kappa}_{m,n}^{\mathrm{co}} &= \frac{1}{N^{m+n-1}} \tilde{\boldsymbol{U}}^T \boldsymbol{W}_\theta^{m\,T} \Theta \boldsymbol{W}_\theta^n \tilde{\boldsymbol{U}}.\end{aligned} \tag{35}$$

We extend this previous work by using these expressions to calculate subpopulation motif cumulants for each individual network unit. Since the initial and final $\widetilde{U}$ terms serve to average over the initial and final subpopulation respectively, we define the average subpopulation motif cumulants into each unit (a $N \times k$ matrix) or out of each unit (a $k \times N$matrix) by removing the $\widetilde{U}^T$ or $\widetilde{U}$ term. Denoting these individual motif cumulant matrices $\hat{\cdot}$, we then have

$$\hat{\boldsymbol{\kappa}}_n^{\mathrm{ch}} = \frac{1}{N^{n-1}} \boldsymbol{W}_n^{\theta} \tilde{\boldsymbol{U}},$$
$$\hat{\boldsymbol{\kappa}}_n^{\mathrm{chR}} = \frac{1}{N^{n-1}} \tilde{\boldsymbol{U}}^T \boldsymbol{W}_n^{\theta}. \tag{36}$$

where $\hat{\kappa}_n^{ch}$ represents the n-chain subpopulation motif cumulants terminating in the unit of interest and $\hat{\kappa}_n^{chR}$ represents n-chain subpopulation motif cumulants originating in the unit of interest. (m,n)-divergent and (m,n)-convergent motifs are symmetric so we only require one expression for each $N \times k$ matrix.

$$\hat{\boldsymbol{\kappa}}_{m,n}^{\mathrm{di}} = \frac{1}{N^{m+n-1}} \boldsymbol{W}_m^{\theta} \Theta \boldsymbol{W}_{\theta}^{n\,T} \tilde{\boldsymbol{U}},$$
$$\hat{\boldsymbol{\kappa}}_{m,n}^{\mathrm{co}} = \frac{1}{N^{m+n-1}} \boldsymbol{W}_{\theta}^{m\,T} \Theta \boldsymbol{W}_{\theta}^{n} \tilde{\boldsymbol{U}}. \tag{37}$$

Statistical synaptic motif model for output unit modulation

Cells which project to the output unit receive direct feedback projections, η, from the output unit, therefore we first examined the relationship between the stimulus- or choice-related firing rate modulation and the magnitude of feedback received. The output unit's activity differs systematically on "go" vs. "no-go" trials, implying that this relationship should be particularly strong for choice-related activity. This prediction is also supported by theoretical work examining the connection between connectivity and dynamics in similar networks[58]. By fitting the models

$$R_{\mathrm{st/ch}} \sim \eta + 1 \tag{38}$$

we observed linear correlations between stimulus and choice-related responses and the strength of the feedback connections, η, both for networks including and excluding STDP with choice-related activity being the most strongly correlated ($r^2_{stimulus} \approx 0.4, r^2_{choice} \approx 0.8$). However, the precise relationship between output feedback and single output unit modulation varies from condition to condition (pre-STDP, post-STDP, EE only, IE only) and network to network demonstrating that this relationship is modulated by changes to the recurrent connectivity. In other words, the coefficients for this model (slope and y-

intercept) were not universal and varied across and within conditions. In other words, feedback magnitude can predict which units will become the most stimulus or choice modulated within a network, but the factors that determine the precise relationship between the feedback strength and a unit's response profile are unexplained.

Our goal then was to explain the observed relationship between firing rate modulation and feedback strength using only the prevalence of individual subpopulation motif cumulants in a universal model which applies across conditions. Such a model provides a detailed understanding of the synaptic features relevant for generating a particular response profile as it accounts for the variability using coefficients tied to each synaptic motif cumulant that apply across conditions. To accomplish this we included these motifs as linear regressors in place of the unknown coefficients from the previous model (equation 37). Summarizing the individual subpopulation motif cumulants of order n_s (i.e., 1 is monosynaptic, 2 is disynaptic, etc.) from subpopulation q (target input units, nontarget input units, output units, inhibitory units) as

$$\kappa_{n_s\,q} = \hat{\kappa}^{\mathrm{ch}}_{n\,q} + \hat{\kappa}^{\mathrm{chR}}_{n\,q} + \sum_{\substack{n=m=1 \\ m+n=n_s}} \left(\hat{\kappa}^{\mathrm{di}}_{m,n\,q} + \hat{\kappa}^{\mathrm{co}}_{n\,q}\right). \tag{39}$$

We can then write a model for the stimulus and choice modulation including all subpopulations q and cumulants up to order n_s as

$$R_{\mathrm{st/ch}} \sim \left(\sum_{n=1}^{n_s}\sum_{q}\kappa_{n\,q}\right)\eta + \left(\sum_{n=1}^{n_s}\sum_{q}\kappa_{n\,q}\right). \tag{40}$$

Note that in this model each synaptic motif cumulant receives two coefficients: one for the interaction with the output feedback and one for feedback independent effect. All coefficients are universal in that they are fit across network conditions meaning that they quantify the contribution of each motif to stimulus and choice modulation under all observed network conditions. This model was fit using the Python statsmodels package with LASSO regularization to remove regressors with small coefficients. All

calculations were done with $n_s = 2$ (monosynaptic and disyanptic motif cumulants) because higher-order models showed marginal improvement as measured by the AIC criterion. Once statistically significant coefficients were identified (p <0.01 Bonferroni-correction for number of motif cumulants included), we refit the model with only significant motif cumulants.

To calculate the overall contribution of a particular motif κ to either stimulus or choice modulation for a unit, $R_{st/ch}$, from Eqn. 39 let β_κ^η be the η interaction coefficient for cumulant κ and β_κ be the independent coefficient and define the contribution from each cumulant as

$$r_{\mathrm{st/ch}}(\kappa) = (\beta_\kappa^\eta \eta + \beta_\kappa)\,\kappa \tag{41}$$

such that

$$R_{\mathrm{st/ch}} = \sum_\kappa r_{\mathrm{st/ch}}(\kappa). \tag{42}$$

Given the definition of overall firing rate modulation, R, as $R^2 = R_{st}^2 + R_{ch}^2$ we can decompose it into a linear combination of motif cumulant related terms

$$\begin{aligned} R^2 &= R_{\mathrm{st}}^2 + R_{\mathrm{ch}}^2 \\ &= R_{\mathrm{st}} \left(\sum_\kappa r_{\mathrm{st}}(\kappa) \right) + R_{\mathrm{ch}} \left(\sum_\kappa r_{\mathrm{ch}}(\kappa) \right) \\ &= \sum_\kappa \left(R_{\mathrm{st}} r_{\mathrm{st}}(\kappa) + R_{\mathrm{ch}} r_{\mathrm{ch}}(\kappa) \right) \\ &\equiv \sum_\kappa r^2(\kappa) \end{aligned} \tag{43}$$

where the contribution of any given cumulant to the overall modulation squared is defined as

$$r^2(\kappa) \equiv R_{\mathrm{st}} r_{\mathrm{st}}(\kappa) + R_{\mathrm{ch}} r_{\mathrm{ch}}(\kappa). \tag{44}$$

Spiking simulations using measured synaptic inputs

For two units, cell-attached spikes were first recorded and then after breaking into the cell, EPSCs and IPSCs were also measured for the same neuron. For all other units, spiking activity was simulated trial-

by-trial using the measured excitatory and inhibitory post-synaptic currents (E/IPSCs) via a integrate-and-fire point neuron simulation based on previously described methods[6]. The simulation relies on conductance-based dynamics

$$\tau_m \frac{\mathrm{d}V}{\mathrm{d}t} = (V_r - V) + g_E(t)R_m(V_E - V) + g_I(t)R_m(V_I - V) \tag{45}$$

where V is the membrane voltage, V_r is the resting membrane potential, and $\tau_m = R_m C_m$ is the membrane time constant. $V_{E/I}$ and $g_{E/I}(t)$ are the excitatory/inhibitory reversal potentials and time based conductances, respectively. Spiking threshold was set to $V_{th} = -40\,mV$ and simulated neurons were constrained with a refractory period of $5\,ms$.

Model parameters for individual neurons were derived using a voltage pulse of $\Delta V = 10\,mV$, the membrane capacitance (C), input resistance (R_i), and series resistance (R_s) were calculated from the current, $I(t)$ via

$$\begin{aligned} R_s &= \frac{\Delta V}{I(0)}, \\ R_i &= \frac{\Delta V}{I(\infty)}, \\ C &= \frac{\tau}{R_i - R_s}, \end{aligned} \tag{46}$$

where $I(0)$ is the initial current on pulse onset, $I(\infty)$ is the asymptotic current, and τ is the measured exponential decay constant for $I(t)$.

PSC time courses in individual trials were fit to a standardized parametric form incorporating the maximum current (I_{max}), the time of max current (t_{max}), the rise time to half max (t_{rise}), and the fall time to half max (t_{fall})

$$I(t) = I_{\max}\left(1 - 2^{\frac{-(t-t_{\max}-t_0)}{t_{\text{rise}}}}\right)\left(2^{\frac{-(t-t_{\max}-t_0)}{t_{\text{fall}}}}\right) \quad (47)$$

where $t_0 = t_{rise}\left(\frac{t_{fall}}{t_{rise}+1}\right)$ is an offset requires so that $I(t_{max}) = I_{max}$. PSCs were then converted to the conductances, $g_{E/I}(t)$ used in the simulation by calculating the synaptic currents $I_{syn}(t) = I(t)\frac{R_{in}+R_s}{R_s}$ and converting to conductance using the holding potential, V_h, $g(t) = \frac{I_{syn}(t)}{V_h - V_r}$. The conductance terms used in the simulation also incorporated noise set to 10% of the average conductance values.

To generate spiking activity for an individual trial, one EPSC and one IPSC measured during stimulus presentation were randomly selected and used to calculate conductance dynamics for simulation. The process was repeated for am EPSC and IPSC taken in the 200 ms prior to stimulus onset to determine baseline activity. This process was repeated 2000 trials and the firing rate modulation during the stimulus period was calculated as the average difference between stimulus-evoked and baseline firing rate.

Predicting in vivo firing rate modulation using motifs from spiking RNN

To validate our spiking RNN model, we applied the statistical synaptic motif model to predict the response profiles of cortical neurons recorded from behaving animals using only measured average synaptic inputs. Because the statistical model was trained on a diversity of network structures and response profile distributions, the coefficients of our model are applicable even in cases where the overall firing statistics differ from that of the full post-STDP network.

To apply our statistical motif model to predict the stimulus-related firing rate modulation of *in vivo* neurons, we first fit a simplified version of our statistical model that only included monosynaptic inputs

from two subpopulations (inhibition and excitation) using data from RNNs under all STDP conditions. This model predicts the stimulus modulation of an individual RNN unit from its average excitatory and inhibitory synaptic inputs. We then used this model to predict the stimulus modulation of a cell *in vivo* by dividing the average excitatory/inhibitory conductance values by the average number of inputs from each cell type in our model (30 for excitation and 10 for inhibition) as a proxy for that cell's average monosynaptic excitatory and inhibitory inputs. These values were then fed to our statistical model to predict the firing rate modulation.

To ensure that these predictions were non-trivial we ran two controls. The 'rescaled conductance' control preserved the average conductance values for each cell but re-ran simulations using trial-by-trial peak conductance that had beet rescaled relative to the mean conductance value for that cell (trial-by-trial peaks were brought closer to the mean by a factor of 2). This control preserved the RNN-derived predictions while altering the dynamics which produced the simulated results. The 'shuffled RNN control' randomly shuffled all unit's modulation values in our RNN before fitting the synaptic motif model to generate the coefficients predicted by chance. These coefficients were then used in place of the true coefficients to determine if the RNN-derived predictions could be explained by chance.

Code availability

All simulation code for this paper is available on Github at albannalab/InsanallyAlbanna2022.

References

1. Musall, S., Kaufman, M. T., Juavinett, A. L., Gluf, S. & Churchland, A. K. Single-trial neural dynamics are dominated by richly varied movements. *Nat. Neurosci.* **22**, 1677–1686 (2019).
2. Goard, M. J., Pho, G. N., Woodson, J. & Sur, M. Distinct roles of visual, parietal, and frontal motor cortices in memory-guided sensorimotor decisions. *eLife* **5**, (2016).
3. Osako, Y. *et al.* Contribution of non-sensory neurons in visual cortical areas to visually guided decisions in the rat. *Curr. Biol.* **31**, 2757-2769.e6 (2021).
4. Guitchounts, G., Masís, J., Wolff, S. B. E. & Cox, D. Encoding of 3D Head Orienting Movements in the Primary Visual Cortex. *Neuron* **108**, 512-525.e4 (2020).
5. Rodgers, C. C. & DeWeese, M. R. Neural correlates of task switching in prefrontal cortex and primary auditory cortex in a novel stimulus selection task for rodents. *Neuron* **82**, 1157–1170 (2014).
6. Kuchibhotla, K. V. *et al.* Parallel processing by cortical inhibition enables context-dependent behavior. *Nat. Neurosci.* **20**, 62–71 (2017).
7. Francis, N. A. *et al.* Small Networks Encode Decision-Making in Primary Auditory Cortex. *Neuron* **97**, 885-897.e6 (2018).
8. Insanally, M. N., Köver, H., Kim, H. & Bao, S. Feature-dependent sensitive periods in the development of complex sound representation. *J. Neurosci. Off. J. Soc. Neurosci.* **29**, 5456–5462 (2009).
9. Otazu, G. H., Tai, L.-H., Yang, Y. & Zador, A. M. Engaging in an auditory task suppresses responses in auditory cortex. *Nat. Neurosci.* **12**, 646–654 (2009).
10. Schneider, D. M., Nelson, A. & Mooney, R. A synaptic and circuit basis for corollary discharge in the auditory cortex. *Nature* **513**, 189–194 (2014).

11. Schneider, D. M., Sundararajan, J. & Mooney, R. A cortical filter that learns to suppress the acoustic consequences of movement. *Nature* **561**, 391–395 (2018).
12. Rodgers, C. C. *et al.* Sensorimotor strategies and neuronal representations for shape discrimination. *Neuron* **109**, 2308-2325.e10 (2021).
13. Raposo, D., Kaufman, M. T. & Churchland, A. K. A category-free neural population supports evolving demands during decision-making. *Nat. Neurosci.* **17**, 1784–1792 (2014).
14. Scott, B. B. *et al.* Fronto-parietal Cortical Circuits Encode Accumulated Evidence with a Diversity of Timescales. *Neuron* **95**, 385-398.e5 (2017).
15. Insanally, M. N. *et al.* Spike-timing-dependent ensemble encoding by non-classically responsive cortical neurons. *eLife* **8**, e42409 (2019).
16. Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).
17. Rigotti, M. *et al.* The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–590 (2013).
18. Leavitt, M. L., Pieper, F., Sachs, A. J. & Martinez-Trujillo, J. C. Correlated variability modifies working memory fidelity in primate prefrontal neuronal ensembles. *Proc. Natl. Acad. Sci.* **114**, E2494–E2503 (2017).
19. Carcea, I. *et al.* Oxytocin neurons enable social transmission of maternal behaviour. *Nature* **596**, 553–557 (2021).
20. Kohl, J. *et al.* Functional circuit architecture underlying parental behaviour. *Nature* **556**, 326–331 (2018).
21. Karigo, T. *et al.* Distinct hypothalamic control of same- and opposite-sex mounting behaviour in mice. *Nature* **589**, 258–263 (2021).

22. Meshulam, L., Gauthier, J. L., Brody, C. D., Tank, D. W. & Bialek, W. Collective Behavior of Place and Non-place Neurons in the Hippocampal Network. *Neuron* **96**, 1178-1191.e4 (2017).
23. Gauthier, J. L. & Tank, D. W. A Dedicated Population for Reward Coding in the Hippocampus. *Neuron* **99**, 179-193.e7 (2018).
24. Fenton, A. A. & Muller, R. U. Place cell discharge is extremely variable during individual passes of the rat through the firing field. *Proc. Natl. Acad. Sci.* **95**, 3182–3187 (1998).
25. Engelhard, B. *et al.* Specialized coding of sensory, motor and cognitive variables in VTA dopamine neurons. *Nature* **570**, 509–513 (2019).
26. Choi, J. Y. *et al.* A Comparison of Dopaminergic and Cholinergic Populations Reveals Unique Contributions of VTA Dopamine Neurons to Short-Term Memory. *Cell Rep.* **33**, 108492 (2020).
27. Guo, L., Weems, J. T., Walker, W. I., Levichev, A. & Jaramillo, S. Choice-Selective Neurons in the Auditory Cortex and in Its Striatal Target Encode Reward Expectation. *J. Neurosci. Off. J. Soc. Neurosci.* **39**, 3687–3697 (2019).
28. Jaramillo, S. & Zador, A. M. The auditory cortex mediates the perceptual effects of acoustic temporal expectation. *Nat. Neurosci.* **14**, 246–251 (2010).
29. Parthasarathy, A. *et al.* Mixed selectivity morphs population codes in prefrontal cortex. *Nat. Neurosci.* **20**, 1770–1779 (2017).
30. Reddy, L. *et al.* Theta-phase dependent neuronal coding during sequence learning in human single neurons. *Nat. Commun.* **12**, 4839 (2021).
31. Leavitt, M. L. & Morcos, A. Selectivity considered harmful: evaluating the causal impact of class selectivity in DNNs. *ArXiv200301262 Cs Q-Bio Stat* (2020).
32. Litwin-Kumar, A. & Doiron, B. Formation and maintenance of neuronal assemblies through synaptic plasticity. *Nat. Commun.* **5**, 5319 (2014).

33. Zenke, F., Agnes, E. J. & Gerstner, W. Diverse synaptic plasticity mechanisms orchestrated to form and retrieve memories in spiking neural networks. *Nat. Commun.* **6**, 6922 (2015).

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

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

36. Bi, G. Q. & Poo, M. M. Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *J. Neurosci. Off. J. Soc. Neurosci.* **18**, 10464–10472 (1998).

37. Song, S., Miller, K. D. & Abbott, L. F. Competitive Hebbian learning through spike-timing-dependent synaptic plasticity. *Nat. Neurosci.* **3**, 919–926 (2000).

38. D'amour, J. A. & Froemke, R. C. Inhibitory and excitatory spike-timing-dependent plasticity in the auditory cortex. *Neuron* **86**, 514–528 (2015).

39. Vogels, T. P., Sprekeler, H., Zenke, F., Clopath, C. & Gerstner, W. Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks. *Science* **334**, 1569–1573 (2011).

40. Field, R. E. *et al.* Heterosynaptic Plasticity Determines the Set Point for Cortical Excitatory-Inhibitory Balance. *Neuron* **106**, 842-854.e4 (2020).

41. Mongillo, G., Rumpel, S. & Loewenstein, Y. Inhibitory connectivity defines the realm of excitatory plasticity. *Nat. Neurosci.* **21**, 1463–1470 (2018).

42. Song, S., Sjöström, P. J., Reigl, M., Nelson, S. & Chklovskii, D. B. Highly Nonrandom Features of Synaptic Connectivity in Local Cortical Circuits. *PLOS Biol.* **3**, e68 (2005).

43. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998).

44. Perin, R., Berger, T. K. & Markram, H. A synaptic organizing principle for cortical neuronal groups. *Proc. Natl. Acad. Sci.* **108**, 5419–5424 (2011).

45. Hu, Y., Trousdale, J., Josić, K. & Shea-Brown, E. Motif statistics and spike correlations in neuronal networks. *J. Stat. Mech. Theory Exp.* **2013**, P03012 (2013).

46. Recanatesi, S., Ocker, G. K., Buice, M. A. & Shea-Brown, E. Dimensionality in recurrent spiking networks: Global trends in activity and local origins in connectivity. *PLoS Comput. Biol.* **15**, e1006446 (2019).

47. Lee, J. H., Delbruck, T. & Pfeiffer, M. Training Deep Spiking Neural Networks Using Backpropagation. *Front. Neurosci.* **10**, (2016).

48. Lillicrap, T. P., Cownden, D., Tweed, D. B. & Akerman, C. J. Random synaptic feedback weights support error backpropagation for deep learning. *Nat. Commun.* **7**, 13276 (2016).

49. Perez-Nieves, N. & Goodman, D. F. M. Sparse Spiking Gradient Descent. 14.

50. Rajan, K., Harvey, C. D. & Tank, D. W. Recurrent Network Models of Sequence Generation and Memory. *Neuron* **90**, 128–142 (2016).

51. Perez-Nieves, N., Leung, V. C. H., Dragotti, P. L. & Goodman, D. F. M. Neural heterogeneity promotes robust learning. *Nat. Commun.* **12**, 5791 (2021).

52. Carcea, I., Insanally, M. N. & Froemke, R. C. Dynamics of auditory cortical activity during behavioural engagement and auditory perception. *Nat. Commun.* **8**, 14412 (2017).

53. Froemke, R. C. *et al.* Long-term modification of cortical synapses improves sensory perception. *Nat. Neurosci.* **16**, 79–88 (2013).

54. Song, S. & Abbott, L. F. Cortical Development and Remapping through Spike Timing-Dependent Plasticity. *Neuron* **32**, 339–350 (2001).

55. Akil, A. E., Rosenbaum, R. & Josić, K. Balanced networks under spike-time dependent plasticity. *PLOS Comput. Biol.* **17**, e1008958 (2021).

56. Brette, R. *et al.* Simulation of networks of spiking neurons: A review of tools and strategies. *J. Comput. Neurosci.* **23**, 349–398 (2007).

57. Engelken, R. Chaotic Neural Circuit Dynamics. (Theoretical and Computational Neuroscience of the Georg-August University School of Science, 2017).

58. Mastrogiuseppe, F. & Ostojic, S. Linking Connectivity, Dynamics, and Computations in Low-Rank Recurrent Neural Networks. *Neuron* **99**, 609-623.e29 (2018).

Acknowledgements: We thank Larry Abbott, Tim Vogels, and David Sussillo for comments and discussions, Silvana Valtcheva for technical guidance with the whole-cell recordings, and Madeline Albanese for assisting with mouse behavioral training. This work was funded by the National Institutes of Health (grant number R00-DC015543 to M.N.I., R01-DC012557 to R.C.F., P01-NS074972 to R.C.F., and U19-NS107616 to R.C.F.), and a NARSAD Young Investigators Award to M.N.I.

Author Contributions: M.N.I., K.K., S.F. and T.G. collected the data. M.N.I, B.F.A., and J.T. performed all model simulations, K.R. and B.D. verified the model, and M.N.I., B.F.A., R.C.F. designed the study. M.N.I., B.F.A., and R.C.F. wrote the paper.

Author Information: The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to mni@pitt.edu or robert.froemke@med.nyu.edu

Figure Legends

Figure 1. Diverse single-unit responses measured in rodent auditory cortex during behavior. a, Schematic of behavior and extracellular tetrode recordings from auditory cortex of rat performing frequency recognition go/no-go task. **b**, Asymptotic behavioral performance for all rats (d' = 2.8±0.1, $p<10^{-4}$, N = 15, Wilcoxon two-sided test). **c**, Example single-unit recording from rat auditory cortex during behavior. **d**, Rasters and peri-stimulus time histograms (PSTHs) for four cortical neurons exemplifying the range from non-classically responsive (red, NCR) to classically responsive (gray, CR). Lines in PSTH, mean firing rate; shading, S.E.M. Horizontal bar, tone duration. **e**, Summary of firing rate modulation for all cortical neurons recorded during behavior (n=103). Outlined circles, units from **d**. Median firing rate modulation = 0.78 spikes/s (inter-quartile range 0.47 – 1.50 spikes/s). **f**, Cell-attached recordings from auditory cortex of mouse performing frequency recognition go/no-go task. **g**, Asymptotic behavioral performance for all mice (d' = 2.45±0.11, p = 0.016, N = 7 mice, Wilcoxon two-sided test). **h**, Example cell-attached recording from mouse auditory cortex during behavior. **i**, Rasters and PSTHs for four example recordings. **j**, Firing rate modulation for all cell-attached recordings (n=26) from mouse auditory cortex during behavior (median firing rate modulation = 2.26, interquartile range = 1.73 - 3.61 spikes/s). **k**, Diagram of relationship between local synaptic structure, synaptic inputs, spiking outputs, and behavior.

Figure 2. A spiking RNN model incorporating STDP rules recapitulating *in vivo* cortical neuronal dynamics. **a,** Schematic of spiking RNN model trained to complete go/no-go stimulus recognition task. Networks consisted of 80% excitatory and 20% inhibitory units. 25% of excitatory units received direct current as the stimulus (1 target stimulus, 6 non-target stimuli) and remaining 75% were output units which project to the readout node (maroon) and received feedback from readout node. Weights to the

readout node trained were via FORCE. Excitatory-to-excitatory and inhibitory-to-excitatory synapses were modified by separate STDP mechanisms. **b,** Example network outputs on 'go' trial (in response to 'target' stimulus) and 'no-go' trial (in response to 'non-target' stimulus). White, pre-trial baseline; gray, stimulus period; green, choice period. **c,** Asymptotic task performance (d' = 4.6±0.1, p=0.0078, N = 8 networks, Wilcoxon two-sided test) **d,** Top left, example voltage traces from two trials of a classically responsive unit. Top right, corresponding spike rasters across trials and PSTHs to target (red) and non-target stimuli (blue). Bottom left, two example voltage traces from a non-classically responsive unit. Bottom right, corresponding spike rasters across trials and PSTHs. **e,** Cumulative distribution of single-unit firing rate modulation for spiking RNN model (dotted) and experimental data (solid). Small values of the firing rate modulation correspond to non-classical response (NCR) profiles; high values correspond to classical response (CR) profiles. Difference between simulated and experimental distributions were not statistically significant (p = 0.27, Kolmogorov-Smirnov test). **f,** Decoding performance for single units in one network (n=1000) for classically responsive units (CR, grey, left) and non-classically responsive units (NCR, red, right). Circle and line represent median and interquartile range respectively. **g,** Probability density of firing rate modulation for individual units. Gray, pre-STDP; purple, post-STDP; dotted light purple, mean-matched control; solid light purple, shuffle control. Summary circles and bars above distributions represent median and interquartile range, respectively (median post-STDP modulation = 1.52 vs. pre-SDTP modulation = 2.25, $p < 10^{-5}$, vs. mean-match modulation = 1.46, $p<10^{-4}$, 52 vs. shuffle modulation = 1.48, p=0.0017, N = 8 networks in all groups, Mann-Whitney two-sided U-test with Bonferroni correction). **h,** Percent of inactive units for pre-STDP, post-STDP, and controls (defined as firing rate <1 spikes/s, all comparisons to post-STDP, p = 0.001, Mann-Whitney two-sided U test with Bonferroni correction). **i,** Task performance for pre-STDP, post-STDP, and controls (Mean shifts relative

to pre-STDP, post-STDP: $\Delta d' = 0.97$, $p = 0.0014$, mean-match: $\Delta d' = 0.44$, $p = 0.04$, shuffle: $\Delta d' = 0.15$, $p = 0.74$, N = 8 networks in all groups, Wilcoxon two-sided test with Bonferroni correction).

Figure 3. Classically and non-classically responsive units contribute to task performance and these response profiles shaped by excitatory and inhibitory STDP. a, Probability density function for output weights from non-classically responsive units (red, NCR) and classically responsive units (grey, CR). N=8 networks, n=4,800 units, $p<10^{-5}$, Levene's test. **b,** Outgoing recurrent synaptic weights from non-classically responsive units (left) and classically responsive units (right). Circles and lines represent median and interquartile range, respectively. Synaptic weights from NCRs were greater overall and when conditioned on target subpopulation (NCR, CR). $p < 10^{-5}$ for all comparisons between NCR and CR, Mann-Whitney U test two-sided with Bonferonni-correction. **c,** Average firing rates of non-classically responsive units (red, NCR) were higher than those of classically responsive cells (grey, CR). Circles and lines represent median and interquartile range, respectively. Median NCR = 19.1 spikes/s vs. median CR = 16.5 spikes/s, $p < 10^{-5}$, Mann-Whitney two-sided U test **d,** Task performance as a function of output units inactivated for non-classically responsive units only (red, NCR), classically responsive units only (grey, CR). Points and bars represent mean and S.E.M., respectively. Inactivating 60 units i.e. 10% most non-classically responsive $\Delta d' = -1.58$ vs. inactivating 10% most classically responsive $\Delta d' = -0.92$, $p < 10^{-4}$; inactivating 300 units i.e. 50% most non-classically responsive $\Delta d' = -2.27$ vs. inactivating 50% most classically responsive $\Delta d' = -2.61$, $p = 0.061$, Mann-Whitney two-sided U test with Bonferroni correction. **e,** Error rates for misses (solid) and false alarms (dotted) as a function of output units inactivated for non-classically responsive units only (red, NCR) and classically responsive units only (grey, CR). Lines and shaded areas represent means and S.E.M. False alarms vs. misses inactivating 50% most classically responsive, $p = 0.26$, inactivating 50% most non-classically responsive, $p = 0.21$, Mann-

Whitney two-sided U test with Bonferroni correction. **f,** mean squared error for readout node activity during choice period as a function of output units inactivated for non-classically responsive units only (red, NCR) and classically responsive units only (grey, CR). Dots and vertical lines represent means and S.E.M. At 10% and 50% inactivation, $p < 10^{-5}$, Mann-Whitney two-sided U test with Bonferroni correction **g,** mean readout node activity during baseline pre-stimulus as a function of output units inactivated for non-classically responsive units only (red, NCR) and classically responsive units only (grey, CR). At 10% and 50% inactivation, $p < 10^{-5}$, Mann-Whitney two-sided U test with Bonferroni correction. **h,** Asymptotic task performance for networks without STDP (pre-STDP), with only inhibitory-to-excitatory STDP (IE only), with only excitatory-to-excitatory STDP (EE only), or all STDP rules (post-STDP). N = 8 matched networks per group each seeded with identical initial weights. Circles represent individual networks and bars represent means. Increase in performance relative to pre-STDP with IE only, $\Delta d' = 1.43 \pm 0.09$, $p = 0.0014$, EE only STDP: $\Delta d' = 0.59 \pm 0.18$, $p = 0.011$, both forms of STDP: $\Delta d' = 1.05 \pm 0.19$, $p = 0.0014$, Wilcoxon two-sided test with Bonferroni correction). **i,** Probability density of firing rate modulation for individual output units for networks without STDP (pre-STDP, light purple solid), with only inhibitory-to-excitatory STDP (IE only, light purple dot dashed), with only excitatory-to-excitatory STDP (EE only, purple dashed), or all STDP rules (post-STDP, purple). Summary circles and bars above distributions represent median and interquartile range, respectively. Median shift in firing rate modulation relative to pre-STDP, Post-STDP: Δmodulation = -0.95 spikes/s, IE only: Δmodulation = -1.13 spikes/s**,** EE only: Δmodulation = -0.79 spikes/s, $p < 10^{-5}$ for all comparisons, Mann-Whitney U test two-sided Bonferroni-correction. Inactive units excluded and percentage shown in **j**.

Figure 4. Specific local synaptic patterns predict response properties of diverse units. a, All possible monosynaptic (first-order) and disynaptic (second-order) motifs. **b,** Network schematic including 4

network subpopulations: target input units, non-target input units, output units, and inhibitory units. **c,** left, observed prevalence of all monosynaptic motifs between individual output units and all subpopulations for non-classically responsive units (NCR, red) and classically responsive units (CR, grey). Bars and lines represent medians and interquartile range, respectively. Right, same as left except for all disynaptic motifs between individual output units and all subpopulations. **d,** Schematic of method for predicting response profile from local synaptic structure around a unit. Local structure was decomposed into a set of synaptic motifs and these motifs were used to predict the modulation of the unit. **e,** Probability density of firing rate modulation for individual output units for networks without STDP (pre-STDP, solid light purple) or all STDP rules (post-STDP, solid dark purple) along with predictions derived from statistical motif model (dotted lines). Small values of the firing rate modulation correspond to non-classical response profiles; high values correspond to classical response profiles. Summary circles and bars above distributions represent median and interquartile range, respectively. **f,** Contributions of individual monosynaptic motifs to single-unit firing rate modulation[2] for classically responsive units (CR, grey) and non-classically responsive units (NCR, red). Circles and bars represent median and interquartile range, respectively.

Figure 5. RNN-derived statistical motif model predicts *in vivo* response profiles. a, Schematic of behavioral task and whole-cell recording set-up during behavior. **b,** An example cell where cell-attached recording was used to first measure spiking outputs before breaking into the cell to record synaptic currents (E/IPSCs). **c,** Example recordings from a non-classically responsive neuron (left) and a classically responsive neuron (right). For non-classically responsive unit recorded spike times are shown; for classically responsive unit simulations are shown. **d,** Simulated firing rate modulation versus actual modulation for 4 neurons where spiking outputs and synaptic inputs were both recorded (Pearson's r = 0.85). **e,** Comparison of cumulative distribution function from cell-attached recordings (black) and leaky-

integrate-and-fire simulation (grey). No significant difference was observed (Kolmogorov-Smirnov test, $p = 0.41$) **f,** Comparison of average synaptic inputs to non-classically responsive neurons (red) versus classically responsive neurons. Bars indicate means and S.E.M. (NCR vs CR, $\Delta Exc = -72\%$, $p = 0.002$, $\Delta Inh = -76\%$, $p = 0.007$, Mann-Whitney two-sided U-test). **g,** recorded/simulated modulation was compared to predictions based on coefficients from the RNN-derived motif statistical model, neurons from **c** circled in black. ($n=12$, mean-squared error = 2.2 spikes/s, Pearson's $r = 0.94$).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- InsanallyAlbanna2022ExtendedDataFigures.pdf