Hierarchical substrates of prediction in visual cortical spiking

Jacob A. Westerberg[1,2,†], Yihan S. Xiong[1,†], Eli Sennesh[1], Hamed Nejat[1], David Ricci[1], Séverine Durand[3], Ben Hardcastle[3], Hannah Cabasco[3], Hannah Belski[3], Ahad Bawany[3], Ryan Gillis[3], Henry Loeffler[3], Carter R. Peene[3], Warren Han[3], Katrina Nguyen[3], Vivian Ha[3], Tye Johnson[3], Conor Grasso[3], Ahrial Young[3], Jackie Swapp[3], Ben Ouellette[3], Shiella Caldejon[3], Ali Williford[3], Peter A. Groblewski[3], Shawn R. Olsen[3], Carly Kiselycznyk[3], Christof Koch[3], Jerome A. Lecoq[3], Alexander Maier[1], and André M. Bastos[1,*]

[1]Department of Psychology, Vanderbilt Brain Institute, Vanderbilt Vision Research Center, Vanderbilt University, Nashville, US
[2]Department of Vision and Cognition, Netherlands Institute for Neuroscience, Royal Netherlands Academy of Arts and Sciences, Amsterdam, The Netherlands
[3]Allen Institute for Neural Dynamics, Seattle, US
[*]Corresponding author. Email: andre.bastos@vanderbilt.edu
[†]These authors contributed equally to this work

Predictive processing models have recently flourished in neuroscience[1–9]. Feedforward and feedback modulation are at the heart of these hierarchical predictive processing models. Previous experimental studies using fMRI, EEG/MEG, and LFP[1,10,11] could not reliably resolve feedback modulation from local computations and feedforward outputs. Here, using open-science[9], multi-species, multi-area, high-density[12], laminar neurophysiology[13], we empirically test whether predictive processing is a key component shaping sensation. To isolate sensory information processing and eliminate motor/reward confounders[1,10,11], we use a no-report task. Our task leveraged so-called global oddballs (GO) as unpredictable, deviant stimuli that circumvent low-level adaptation. We examined their responses relative to local oddballs (LO) that we habituated into highly predictable priors. Four surprising findings in this dataset challenge many existing predictive processing models. First, passively evoked GO responses were exclusive to higher-order, more cognitive areas rather than early-to-mid sensory cortex. Second, interneuron-targeted optogenetics revealed no evidence for a subtractive mechanism in both primates and mice. Third, highly predictable LO responses dominated in over 50% of all neurons, including in higher-order cortex which should have anticipated them, indicating limited evidence for predictive suppression. Lastly, prediction errors followed a feedback, rather than a feedforward signature. These results reveal circuit dynamics that govern the shaping of sensory processing by prediction, which will motivate new, neurally-constrained predictive processing models.

Introduction

Predictive Processing (PP) models state that brains have evolved to model the sensory statistical regularities in the world[2,14,15]. Deviations from these internal predictions are neuronally signaled with prediction error responses[16]. These models (see Extended Data Table 1 and Extended Data Fig. 1) make specific hypotheses about the neuronal circuitry that implements prediction and prediction error computations.

Sequence-based oddball paradigms have been widely used to test PP models[1]. Subjects are exposed to a repeated pattern of sensory inputs to implicitly learn the statistical regularities of hidden rules governing the sensory inputs. Violations to the rule are used to test for neuronal signaling of prediction errors. Studies of brain activity during oddball tasks using fMRI, MEG/EEG, and LFP have been largely consistent with PP models[1,10,11,13,17–21] and have documented neuronal signatures of prediction errors both in higher-order areas (e.g., prefrontal cortex) as well as sensory areas, including primary and secondary auditory and visual cortex[13]. However, because these neuronal recording techniques cannot resolve feedback modulation from local computations and feedforward output[22], they could not test key hypotheses made by the PP models.

In this work we test three of the main hypotheses made by PP models. The first hypothesis is that predictions are generated in higher-order areas of the brain and feed back to lower-order areas, where they are compared with and subtracted from sensory inputs (Fig. 1, Hypothesis 1). In the second hypothesis, prediction errors are hypothesized to flow in the opposite direction. They feed forward up the hierarchy to update internal models to make better predictions (Fig. 1, Hypothesis 2). The subtractive feedback predictions are thought to be implemented via the activity of local inhibitory interneurons[2,4,23] (Fig. 1, Hypothesis 1). Predictable stimuli are uninformative, so they are hypothesized to drive less overall neuronal activity to save energy[3]. By contrast, surprising/unpredictable stimuli evoke prediction errors which enhance neuronal activity. In the third hypothesis, PP models stated that brains construct complex models that are compared to sensory inputs at multiple levels of cortical processing (Fig. 1, Hypothesis 3). Computational instantiations of PP models have proposed

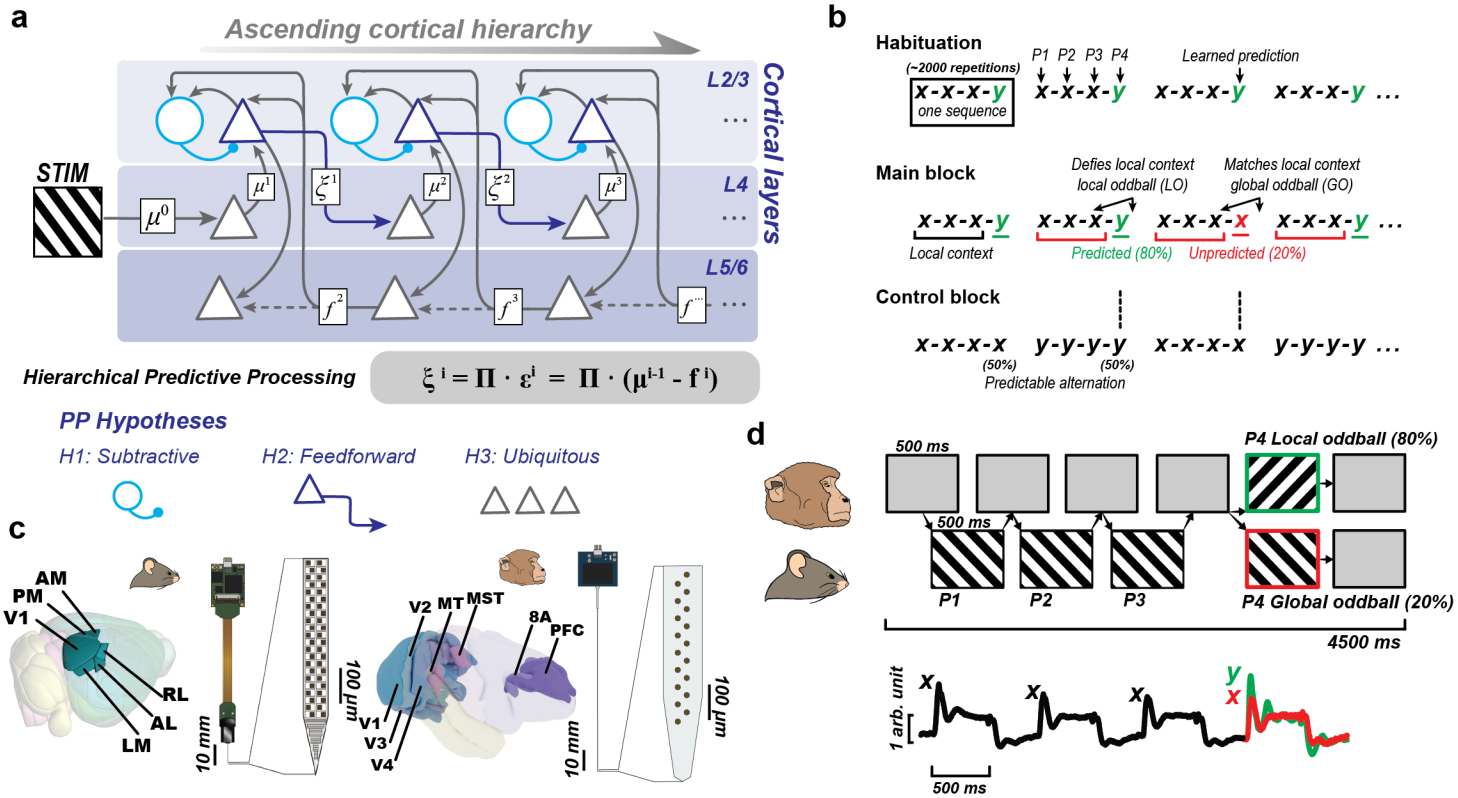

Fig. 1 | Predictive Processing (PP) hypotheses, Local/global oddball paradigm, and experimental setup. **a,** Schematic for sensory processing via hierarchical predictive processing. Stimulus information enters the cortical hierarchy through layer 4, and prediction error is computed in superficial layers 2/3, with inhibitory cell involvement in the subtractive process (H1). Prediction error is then fed forward (H2) via superficial layers to higher-order areas. Higher-order areas provide a feedback prediction from deep layers. Error signals emerge early and propagate throughout cortical hierarchy (H3) **b,** Visual stimuli were presented in 4-stimulus sequences consisting of 2-oriented, drifting gratings symbolized as 'x' or 'y'. Animals were habituated to the x-x-x-y sequence (counterbalanced over sessions/animals with y-y-y-x). In the main block of a recording session, animals were presented with 80% x-x-x-y and 20% x-x-x-x (global oddballs). Control blocks consisted of x-x-x-x (control for global oddball, 50% of trials) and y-y-y-y (control for local oddball, 50% of trials). **c,** Neuropixels were introduced into 6 visual cortical regions in all mice (V1, LM, RL, AL, PM, AM). 128-channel or 32-channel laminar probes were introduced into 8 cortical regions in 2 monkeys (V1, V2, V3, V4, MT, MST, 8A, PFC). **d,** Monkeys viewed sequences of drifting grating stimuli while fixating centrally. Mice viewed sequences of full-screen drifting grating stimuli. Trace shows example population average response to the 4-stimulus sequences with the GO in red and the LO in green.

prediction error computations to occur both in early sensory as well as higher-order levels of neuronal circuits, even for high-level predictions[11,24,25].

So far, studies of spiking activity have been more equivocal, with different results depending on species, paradigm, cortical area, and methodology[4,13,21,26–30]. Importantly, no study has yet used large-scale high-density neuronal electrophysiological recording methods in multiple connected brain regions to test PP models. Brain recordings with this level of detail are necessary to experimentally test PP because these models make use of multiple stages of cortical processing, with distinct cell types, cortical layers, areas, and directions of feedforward/feedback processing contributing to the hypothesized computations.

We tested the above hypotheses regarding PP using Multi-Area, high-Density, Laminar-resolved Neurophysiology (MaDeLaNe) recordings of thousands of spiking neurons in mice and monkeys throughout the visual cortex (areas V1, LM, RL, AL, PM, AM in mice and areas V1, V2, V3, V4, MT, MST in monkeys) and prefrontal cortex (areas 8A and lateral PFC in monkeys). We used a habituation-based no-report variation[21,31] of the global-local oddball task[1], thereby avoiding motor and reward related confounds on sensory responses. This paradigm allowed us to disentangle short-term stimulus repetition from prediction. We found four surprising findings which challenge many aspects of current PP models (Extended Data Table 1, Extended Data Fig. 1). First, global oddballs evoked neuronal spiking responses consistent with a prediction error signal, but these responses were restricted to higher-order cortical areas in both species. Second, we used cell-type specific optogenetics in both species to observe whether inhibitory interneurons performed predictive inhibition as proposed. Global oddballs did not modulate the inhibitory interneuron populations we studied. Third, highly expected local oddballs did not evoke a reduced neuronal response compared to the same sequence when it was contextually deviant. Fourth, global oddballs evoked a laminar- and area-wise pattern of activity more consistent with feedback, rather than feedforward processing.

Isolating sensation and establishing priors

We first extensively habituated animals to the local oddball x-x-x-y stimulus sequence for 2000-3000 trials over several days (Fig. 1b and Extended Data Fig. 2). The "x" and "y" stimuli within a sequence denote drifting gratings at 135° or 45° from horizontal and were counterbalanced across animals. The unpredictable global oddball x-x-x-x stimulus sequence was never presented during habituation and introduced only on neural recording sessions (Fig. 1c), on 20% of trials in the main block (Fig. 1b). This long habituation period ensured that animals could learn to predict the x-x-x-y (local oddball) stimulus sequence. Pupillary responses suggested that animals recognized the violation of this sequence during unpredicted x-x-x-x global oddballs (Extended Data Fig. 3).

To investigate the neuronal signaling underlying these predictions, we compared responses in the main block to a control block where the x-x-x-x sequence alternated with y-y-y-y and thus the sequence was predictable after the first stimulus (Fig. 1b). To analyze global and local oddballs, we compared spiking responses to the fourth component in the sequence (P4) to the same stimulus in the same sequence position of the control block (vertical dotted lines in Fig. 1b). To account for cross-session firing rate drift which occurs even in well-isolated single units[32], we normalized responses to the third component in the sequence (P3). Our main contrast was therefore P4-P3 in the main block vs. P4-P3 in the control block. For global oddballs, this contrast should reveal prediction errors while controlling for short-term adaptation. For local oddballs, this contrast should reveal the release of adaptation caused by changing the stimulus after a few repetitions. Neurons responded to these sequences with robust firing rate increases in all recorded areas (Fig. 1d, Extended Data Fig. 4, 5). Significant oddball detection was determined by comparing the neuronal spiking response for each oddball type using a nonparametric, cluster-based permutation test[33] (at $P < 0.05$, corrected for multiple comparisons, see *Methods*).

Predictable local oddballs are widely signaled

Because of extensive habituation, local oddballs were non-surprising stimuli and, according to multiple PP models, should be suppressed and explained away (H1). However, we observed that neuronal signaling of local oddballs was ubiquitous (present in all areas and both species, Fig. 2a), signaled early and fed forward up the visual cortical hierarchy (within the first ~150 ms of stimulus processing, Fig. 2b), strong (local oddball responses by area were on average 60-98% above the control stimulus in mice, and 43%-93% in monkeys, Fig. 2c), and involved more than 50% of all recorded neurons (median across areas in mice: 51%, median across areas in monkeys: 65%, Fig 2d). In mice but not monkeys, local oddball signaling gradually increased in strength with ascending cortical hierarchy (Linear correlation between local oddball response and hierarchical area, $R^2 = 0.83$, $P=5.75e^{-5}$, Fig. 2c, left subpanel). In monkeys, the local oddball response was strongest in area MT (ANOVA, $p<0.001$, Fig. 2c, right subpanel). In V1 of both species, local oddball signaling was most prominent amongst neurons in L2/3 neurons (Fig. 2e; for details on layer identification, see *Methods* and Extended Data Fig. 6).

Can these local oddball responses, representing local signal change, be considered a simple form of prediction error signaling? To examine this, we compared the local oddballs in contexts that differed in their relative probability (local oddball sequence probability varied between 100%, 80%, and 12.5%, see Extended Data Fig. 2). If they represent prediction errors, local oddball responses should scale according to their deviance (H1). We found that local oddball signaling did not scale as a function of deviance in most areas (Extended Data Fig. 7, area V4 was an exception). Instead, we observed the opposite pattern just as often, and most prominently in V1 of mice (Extended Data Fig. 7b). The enhanced neuronal responses to more predictable local oddballs in mouse V1 may have been caused by learning-driven potentiation[34]. To summarize, local oddball sequences (x-x-x-y) release most neurons from adaptation and increase excitability independent of prediction, even though animals had been exposed to these sequences thousands of times.

Unpredictable global oddballs do not generate feedforward prediction error

Next, we tested whether highly surprising global oddballs (x-x-x-x) drive neuronal spiking consistent with prediction error signaling. We found that global oddball signaling was restricted to only a few higher-order brain areas (LM, AM, and PM in mice; V3, MT, 8A, and PFC in monkeys, Fig. 3a). At the individual neuron level, although we identified a handful of neurons/areas that responded to global oddballs (Extended Data Fig. 4, 5), the percentage of units signaling global oddballs in each area was sparse (median across areas in mice: 7%, median across areas in monkeys: 8%, Fig. 3b), contradicting H3. Unlike local oddballs, which showed a clear temporal progression of latencies across the hierarchy (Fig. 2c), the latency at which neurons signaled global oddballs did not scale with hierarchy (Fig. 3c).

We next examined global oddballs within L2/3 pyramidal neurons, which according to multiple PP models should transmit prediction error (H2). In mice, we found that putative layer 2/3 pyramidal cells did not signal global oddballs (Extended Data Fig. 8a). In addition, current source density analysis of synaptic activity in L2/3 of mice did not reveal any reliable changes in synaptic activation (Extended Data Fig. 9a, b) during global oddballs. In monkeys, L2/3 spiking activity did not signal global oddballs in area V3 (Extended Data Fig. 8b). There was L2/3 involvement in global oddballs in area MT, but this came 116 ms later than

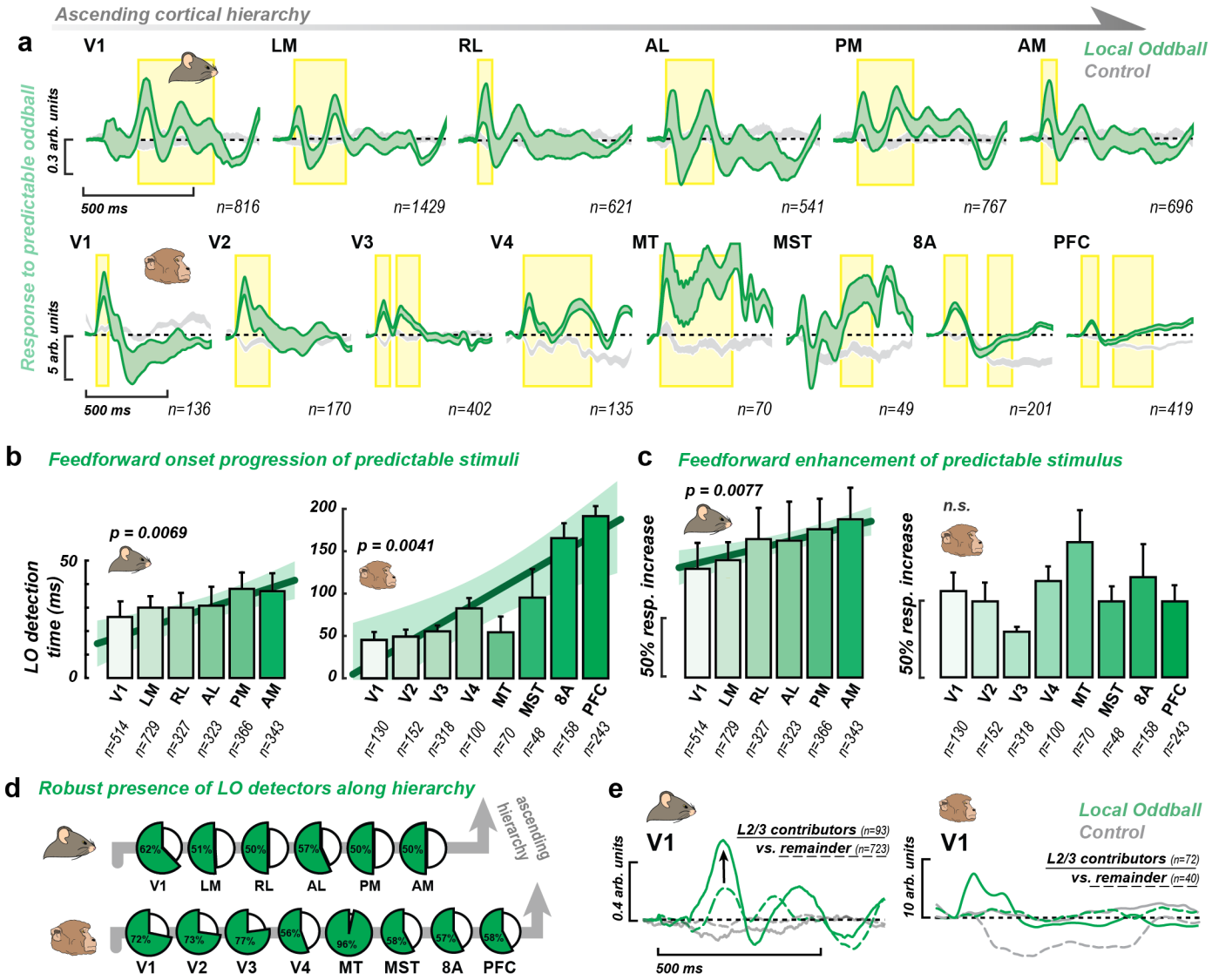

Fig. 2 | Predictable local oddball response is widespread, emerges early, and feeds forward. a, Local oddball detection across cortical areas in mice and monkeys. Bands are 95% confidence intervals across units in an area. Unit responses are normalized by dividing instantaneous firing rate by average firing rate at the single-unit level in mice and via a z-score method for the multiunit data in monkeys (see *Methods*). Green bands are the P4-P3 local oddball in the main block; gray bands are the P4-P3 local oddball in the control block. For non-subtracted neuronal activity, see Extended Data Fig. 4. Yellow highlights reflect significant population local oddball detection periods, P<0.05, corrected using nonparametric, cluster-based permutation tests. Local oddball signaling in mice was accompanied by a stimulus-induced temporal modulation corresponding to the drifting grating presentation rate of 4 Hz (upper subpanel). This oscillatory component was much less prominent in monkeys (lower subpanel). **b,** Average onset time for significant local oddball effect across hierarchy in mice and monkeys. Error bars indicate 2 SEM across units. Linear regression shows significant relationship between hierarchical position and onset timing ($R^2 = 0.83$, P=0.006 in mice; $R^2 = 0.88$, P= 0.004 in monkeys), indicating a feedforward progression. **c,** Percent response increase to local oddball across cortical areas in mice and monkeys. Error bars indicate 2 SEM across units. Local oddball effects show significant feedforward enhancement in mice along the cortical hierarchy. **d,** Percentage of local oddball encoding units along cortical hierarchy in mice and monkeys. Signaling of predictable local oddballs is robust (>50% of all cortical areas) and widespread (observed in all cortical areas) **e,** Local oddball responses are more pronounced in L2/3 in V1 in both mice and monkeys compared to units in other layers, suggesting feedforward signaling.

in area PFC, supporting a feedback[35,36] rather than feedforward putative error computation (Extended Data Fig. 8b).

Without the hypothesized L2/3-specific signal, we investigated alternative sources for global oddball signaling. We limited the analysis to the areas with population global oddball signaling. We grouped areas into higher- and lower-order (higher-order areas AM and PM vs. lower-order area LM in mice; higher-order area PFC vs. lower-order areas V3 and MT in monkeys). We first divided the units into granular (L4) vs. extragranular (L2/3 and L5/6) compartments, which anatomically separate feedforward vs. feedback input[37]. In mice, we observed global oddball signaling only in the extragranular (feedback-associated) layers and later in the lower-order cortex (401 ms post-stimulus, Fig. 3b, left subpanels) than in the higher-order cortex (337 ms post-stimulus). In monkey visual areas V3 and MT, we observed

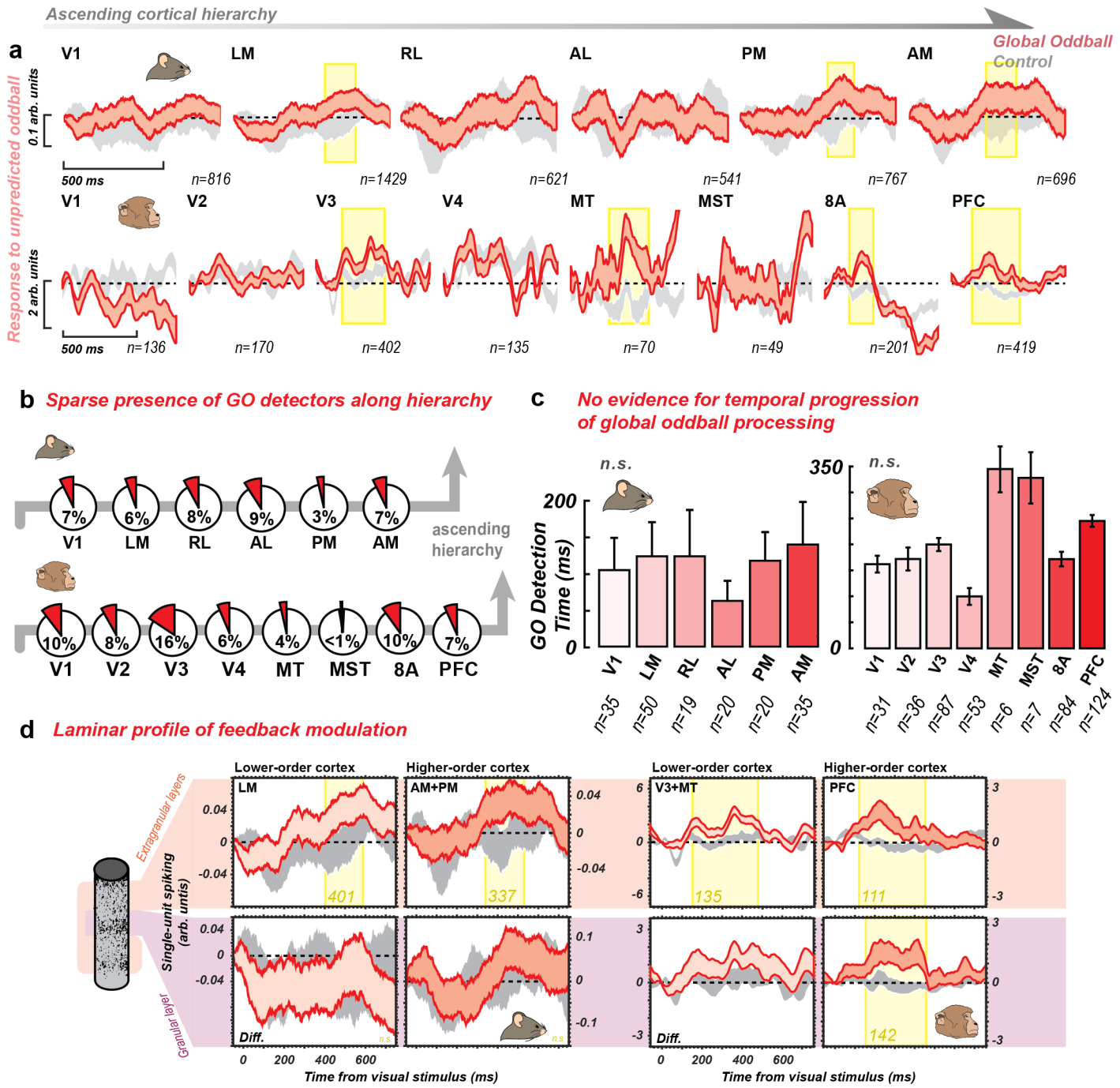

Fig. 3 | Unpredictable global oddball emerges in higher-order cortex and is fed back. a, Global oddball detection across cortical areas in mice and monkeys. Bands are 95% confidence intervals across units in an area. Red bands are P4-P3 in the GLO block; gray bands are P4-P3 in the control block. For non-subtracted neuronal activity, see Extended Data Fig. 4. Yellow highlights reflect periods of significant population global oddball detection, P<0.05, corrected for multiple comparisons using nonparametric, cluster-based permutation tests. **b,** Percentage of significant global oddball encoding units/channels along the cortical hierarchy in mice and monkey. **c,** Onset times of significance for global oddball effect along the cortical hierarchy in mice and monkeys. Linear regressions show no significant change along cortical hierarchy, suggesting no strong feedforward signaling of global oddball prediction error. **d,** Granular versus extra-granular global oddball detection in areas with significant population-level detection in mice and monkeys. Red bands are P4-P3 in the GLO block; gray bands are P4-P3 in the control block. Yellow highlights reflect periods of significant population global oddball detection in the layer grouping, with the latency indicated.

global oddball signaling only in the extragranular (feedback-associated) layers (135 ms post-stimulus), later in time than the extragranular layers of higher-order area PFC (111 ms post-stimulus, Fig. 3b, right subpanels). These latency and laminar effects collectively support prediction error propagation from higher to lower areas of the hierarchy, contradicting H2.

Finally, we also analyzed whether global oddballs were more consistently signaled in the first few trials when they occurred early in the main block, when putative prediction

errors should be strongest. Although area V4 in monkeys and PM in mice exhibited a profile consistent with a putative prediction error signal, areas before or after them in the hierarchy did not (Extended Data Fig. 10), inconsistent with the hypothesized feedforward (H2) or ubiquitous (H3) prediction error signal.

Directed connectivity of global oddballs does not indicate feedforward error propagation

In multiple PP models, prediction errors are sent feedforward (Fig. 1a, H2). Hence, neurons earlier in the hierarchy should increase their drive onto neurons later in the hierarchy to signal prediction error. We tested this by computing Granger causality between spiking activity time courses. Granger causality tests whether neuronal activity in area A can statistically predict activity in area B above and beyond predictions made by activity in area B alone[37]. If so, then area A "Granger-causes" B. By comparing the A-to-B vs. the B-to-A directions of Granger causality[38] where A is below B in the hierarchy, we tested whether local and global oddballs evoked feedforward processing. In mice, compared to pre-stimulus baseline (Extended Data Fig. 11a), the hierarchy became more feedforward-dominated in both early (100-300 ms) and late (300-500 ms) periods of local oddball processing (Extended Data Fig. 11a, b; Wilcoxon rank sum test, P<0.01). This feedforward activity propagated from V1 to the highest levels of the visual hierarchy, area PM. In contrast, during global oddballs, there was no change in feedforward vs. feedback asymmetry compared to baseline (Extended Data Fig. 11b-d, Wilcoxon rank sum test of GC asymmetry values across mice vs. pre-oddball baseline, all comparisons, P>0.01). To summarize, the latency, hierarchical areas of significant signaling, laminar, and Granger analyses collectively contradicted H2; namely, global oddballs did not drive feedforward processing. Instead, the data supported a feedback model for processing prediction errors evoked by global oddballs.

Global oddballs: A release from predictive inhibition?

In multiple PP models, predictions are hypothesized to subtract away predictable stimuli (H1), leaving sensory cortex less excited. We tested this by assessing whether predictions are neuronally instantiated by inhibitory interneuron activity. We used optogenetics to identify Somatostatin- (SST) and Parvalbumin- (PV) expressing interneurons in mice, and a viral vector tool that restricted

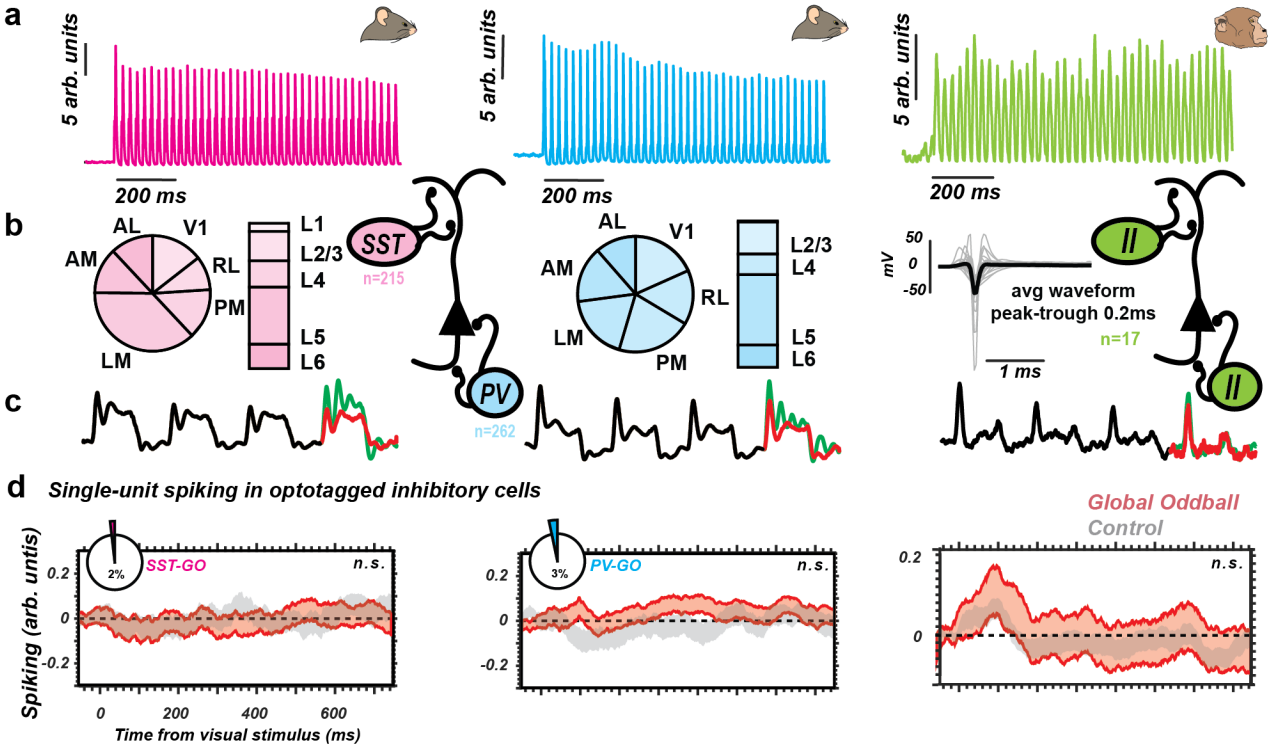

Fig. 4 | Optotagged inhibitory cells in mice and monkeys do not carry out hypothesized predictive coding functions. a, Optical stimulation with 40 Hz sinusoidal modulation of laser power. Tagged SST (magenta, n=215 neurons) and PV (cyan, n=262 neurons) responses are normalized by their average firing rate. In primates, an opsin targeting all GABAergic inhibitory cells was used; responses shown were computed as a weighted mean (green, n=21 neurons), where each cell is weighted by its modulation strength (average firing rate change from baseline, divided by the sum of baseline and laser-timed response). **b,** Relative proportion of each inhibitory interneuron type, which have synapses at largely different sites on pyramidal cells (schematized here), across areas and layers in mice (left and middle). Average waveform shape and spike width in primate data (right) are consistent with that of narrow-spiking interneurons. **c,** Population average spiking response to the oddball sequences for these inhibitory interneurons. P1 to P3 in black, local oddball P4 in green, and global oddball P4 in red. **d,** Lack of global oddball detection in SST and PV cell subpopulations in mice and in pan-inhibitory cell populations in monkeys. Bands are 95% confidence intervals across units. Top, the red band is P4-P3 in the main block; the gray band is P4-P3 in the control block. For P3 and P4 before subtraction, see Extended Data Fig. 4B.

optogenetic expression to inhibitory interneurons in monkeys[39] (Fig. 4a and Extended Data Fig. 12). In both species, we utilized Channelrhodopsin-2 to drive the inhibitory interneurons to fire during light stimulation (see *Methods*). In mice, PV+/SST+ units were identified across all recorded areas and in monkeys, inhibitory interneurons were identified in areas MT/MST, PFC, and 8A (Fig. 4b). In monkeys, opto-tagged neurons had on average narrow waveforms, additional confirmatory evidence[40] that these opto-tagged neurons were inhibitory (Fig. 4b). In both species, these inhibitory cell populations responded to the stimulus sequence and increased their spike rates during P4 of local oddballs (Fig. 4c). However, we did not observe a significant response to global oddballs at the population average of these subpopulations (Fig. 4d; for P3 and P4 components see Extended Data Fig. 4b). This indicates a minor role, if any, for these inhibitory interneuron sub-populations in signaling global oddballs, failing to support H1 and indicating a putative excitatory mechanism for feedback-driven GO modulation.

Discussion

To test the neuronal implementation of PP, we performed an extensive survey of cortical spiking in mice and monkeys observing predictable or unpredictable stimulus sequences. We used MaDeLaNe (Multi-Area, high-Density, Laminar Neurophysiology) recordings[13] and optogenetics to isolate different cortical layers, areas, and neuron types previously hypothesized to play a role in PP[2–4,41]. We tested whether neuronal spiking responses could be described as prediction error signals resulting from a release from predictive suppression (H1), and that triggers feedforward processing (H2) with a ubiquitous signature (H3).

PP models (Extended Data Table 1) have hypothesized that prediction errors constitute a core, feedforward cortical computation[2–4,41]. Our subjects were habituated to a particular sequence (x-x-x-y) for thousands of trials such that its violation (x-x-x-x) should drive robust spike signaling of prediction error. In PP, the habituated (and highly predictable) local oddballs (x-x-x-y) should be suppressed, driving less activity[42]. We found the exact opposite: spiking responses robustly signaled unsurprising local oddballs (engaging over 50% of all recorded neurons) but weakly signaled surprising global oddballs (engaging less than 10% of all recorded neurons). Local oddballs emerged early in the hierarchy at fast latencies in layers 2/3 and fed forward up the cortical hierarchy (Fig. 5a). However, contradicting the interpretation of these signals as prediction errors, local oddball responses did not consistently scale with increased sequence deviance, and did not trigger a release of inhibition from inhibitory interneurons (contradicting H1). Therefore, local oddball detection is most likely a release from neuronal adaptation.

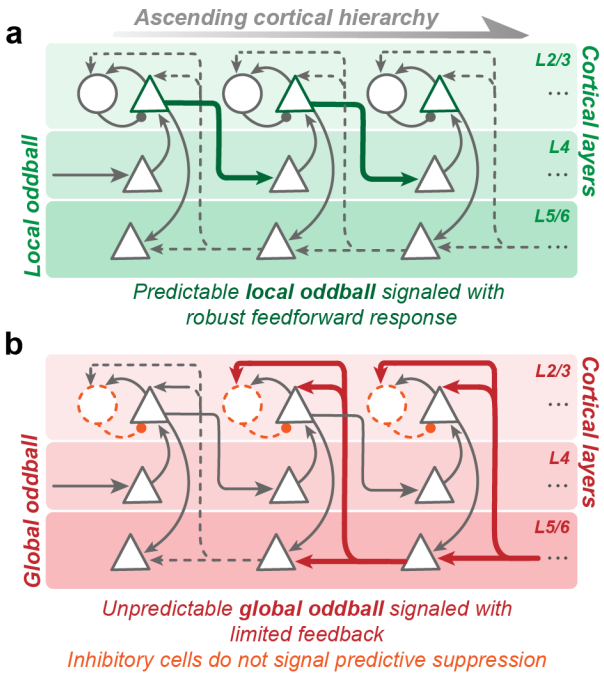

Fig. 5 | Schematic of findings represented in the canonical cortical connections. a, highly predictable local oddball resulted in robust and widespread feedforward response. The timings of onset of significance as well as granger causality results suggest strong feedforward processing. Thick green arrows represent strong cortical activity across layers and widespread across cortical hierarchy. **b,** highly unpredictable global oddball resulted in limited feedback response, particularly in select higher order areas. We did not observe L2/3 prediction error signals in mice and found non-specific L2/3 (compared to L5/6, except for in PFC at top of the hierarchy) prediction error signals in monkeys. Furthermore, optotagged inhibitory cells do not exhibit suppression specific to predictable stimuli.

Multiple PP models hypothesized that L2/3 pyramidal neurons transmit feedforward prediction errors. Global oddballs, which should have triggered this hypothesized feedforward flow, emerged late in the sensory response and in higher-order areas first, then lower-order regions (Fig. 5b). Putative excitatory pyramidal neurons in L2/3 (which project feedforward connections) and their associated current sinks did not signal global oddballs in mice and only signaled weakly in one area (monkey area MT) and later than in PFC. Instead, global oddballs were found to emerge at the population level in the non-granular, feedback-recipient layers. Feedforward neuronal communication assessed by Granger causality amongst spiking neurons was not observed during global oddballs. Therefore, we failed to detect feedforward processing during global oddballs by four independent metrics: hierarchical area, temporal order, Granger causality, and laminar compartment of spiking/transmembrane current flow (Fig. 5b, contradicting H2). Further, global oddballs did not emerge because of a release of inhibition from PV/SST+ interneurons in mice or inhibitory interneurons in monkeys (contradicting H1). Our results speak against the proposed ubiquity of prediction error signaling in sensory cortex (contradicting H3).

Altogether, our results challenge multiple aspects of current PP models[2–4,41].

We found that PP is largely a property of neurons in higher-order cortex (PFC in primates, and late stages of visual processing in both species). These higher-order areas, in which we found population level coding of global oddballs[13,21,31,43], contain neurons with long timescales[44] and therefore may be able to represent the relatively long duration of a sequence and its context, necessary elements to drive PP in this task. In contrast, neurons early in the visual sensory cortex have rigidly-defined receptive field properties refined during brain development[45] with less plasticity in the mature brain. Neurons in sensory cortex also respond to fast changes in the environment and are therefore ill-suited for predictive computations that require flexible re-mapping of responses and contextual processing over longer time scales. Neurons in higher-order areas are more flexible in their response preferences. These higher-order neurons can re-map their activity in real-time based on experience-dependent learning and display mixed-selectivity. This property is necessary for predictive processing because predictions need to be sensitive to the statistical structure of a changing and context-dependent environment[11,22,46]. We hypothesize that mixed selectivity neurons flexibly remap cognitive spaces to signal predictions. Functionally, such neurons could form dynamic ensembles via neuronal oscillations[21,47] to guide sensory processing.

Although our empirical findings failed to support multiple aspects of current PP models (Extended Data Table 1, Extended Fig. 1), some PP models hypothesized circuit elements which were supported by our data. For example, Nejad et al.[6] propose that both deep and superficial layers cooperate for prediction error computations, an effect highly aligned with our laminar results. Higher-order areas participated in PP, although the computation was sparse across neurons. This supports models that have proposed that PFC and higher-order areas are involved in building flexible prediction in real-time[21,31] along sub-spaces of neuronal activity (rather than engaging prediction errors widely in a core computation). Our result on global oddballs being more consistent with feedback (rather than feedforward) processing has yet to be considered in the PP model family, as far as we are aware. In addition, not all aspects of the proposed PP models have been tested. With our full dataset now openly available, this will be a rich resource as a community effort to establish the PP principles and circuitry that are more consistent with MaDeLaNe data. Using novel approaches to model building[48] we suggest that PP models[9] mature to be constrained and grounded in these and other emerging MaDeLaNe neuronal recordings[12,13,49].

Acknowledgments | This work was funded by the US National Institutes of Health (NIH) [grant numbers: R00MH116100 (AMB), U24NS113646 (JAL, CrK)], the Dutch Research Council (NWO) [grant number: VI.Veni.232.110 (JAW)], the International Human Frontier Program Organization (HFSPO) [grant number: LT0001/2023-L], the Vanderbilt Faculty Fellow Award (AMB), and Vanderbilt University Startup Funding (AMB). The Neuropixels dataset was obtained at the Allen Brain Observatory as part of the OpenScope program, which is operated by the Allen Institute, Neural Dynamics program. We thank the OpenScope steering committee for their support, the Allen Institute founder, Paul G. Allen and Karel Svoboda, for their vision, encouragement, and support.

Author contributions | Conceptualization: JAW, AM, AMB; Data curation: JAW, AB, SD, BH, JAL; Formal analysis: JAW, YSX, ES, AMB; Funding acquisition: JAW, JAL, AMB; Investigation (NHP studies): JAW, YSX, HN, AMB; Investigation (rodent studies): SD, BH, HC, HB, HL, WH, KN, VH, TJ, CG, AY, JS, RG, BO, SC, AW, PAG; Project administration: JAW, CaK, JAL, AM, AMB; Software: JAW, AB, CRP; Supervision: SRO, ChK JAL, AM, AMB; Validation: JAW, AB, CRP, JAL; Visualization: JAW, YSX; Writing – original draft: JAW, AM, AMB; Writing – review, and editing: all authors

Competing interests | Authors declare that they have no competing interests.

Supplementary information | Methods included below. Supplementary information includes Supplementary text, Extended Data Table 1, Extended Data Figs. 1-12, and References 50-92.

References

1. Bekinschtein, T. A. *et al.* Neural signature of the conscious processing of auditory regularities. *Proc. Natl. Acad. Sci* **106**, 1672–1677 (2009).
2. Rao, R. P. N. & Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci* **2**, 79–87 (1999).
3. Friston, K. The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci* **11**, 127–138 (2010).
4. Bastos, A. M. *et al.* Canonical Microcircuits for Predictive Coding. *Neuron* **76**, 695–711 (2012).
5. Keller, G. B. & Mrsic-Flogel, T. D. Predictive processing: a canonical cortical computation. *Neuron* **100**, 424–435 (2018).
6. Nejad, K. K., Anastasiades, P., Hertäg, L. & Costa, R. P. Self-supervised predictive learning accounts for cortical layer-specificity. *Nature Communications* **16**, 6178 (2025).
7. Van Derveer, A. B. *et al.* A role for somatostatin-positive interneurons in neuro-oscillatory and information processing deficits in schizophrenia. *Schizophrenia Bulletin* **47**, 1385–1398 (2021).
8. Spratling, M. W. Reconciling predictive coding and biased competition models of cortical function. *Front. Comput. Neurosci* **2, 4**, (2008).
9. Aizenbud, I. *et al.* Neural mechanisms of predictive processing: a collaborative community experiment through the OpenScope program. Preprint at https://doi.org/10.48550/arXiv.2504.09614 (2025).
10. Chao, Z. C., Takaura, K., Wang, L., Fujii, N. & Dehaene, S. Large-scale cortical networks for hierarchical prediction and prediction error in the primate brain. *Neuron* **100**, 1252–1266 (2018).
11. Wacongne, C. *et al.* Evidence for a hierarchy of predictions and prediction errors in human cortex. *Proc. Natl. Acad. Sci* **108**, 20754–20759 (2011).
12. Siegle, J. H. *et al.* Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature* **592**, 86–92 (2021).
13. Gabhart, K. M., Xiong, Y. & Bastos, A. M. Predictive coding: a more cognitive process than we thought? Trends Cogn. *Sci* **0**, (2025).
14. Helmholtz, H. *Handbuch Der Physiologischen Optik. English Translation*. vol. 3 (Dover, New York, 1860).
15. Srinivasan, M. V., Laughlin, S. B. & Dubs, A. Predictive coding: a fresh view of inhibition in the retina. *Proceedings of the Royal Society of London. Series B. Biological Sciences* **216**, 427–459 (1982).
16. Schultz, W. & Dickinson, A. Neuronal coding of prediction errors. *Annual review of neuroscience* **23**, 473–500 (2000).

17. Uhrig, L., Janssen, D., Dehaene, S. & Jarraya, B. Cerebral responses to local and global auditory novelty under general anesthesia. *Neuroimage* **141**, 326–340 (2016).
18. Jiang, Y. *et al.* Constructing the hierarchy of predictive auditory sequences in the marmoset brain. *elife* **11**, e74653 (2022).
19. Chennu, S. *et al.* Expectation and attention in hierarchical auditory prediction. *Journal of Neuroscience* **33**, 11194–11205 (2013).
20. Nourski, K. V. *et al.* Auditory Predictive Coding across Awareness States under Anesthesia: An Intracranial Electrophysiology Study. *J. Neurosci* **38**, 8441–8452 (2018).
21. Xiong, Y. *et al.* Propofol-mediated loss of consciousness disrupts predictive routing and local field phase modulation of neural activity. *Proc. Natl. Acad. Sci* **121, e2315160121**, (2024).
22. Logothetis, N. K., Pauls, J., Augath, M., Trinath, T. & Oeltermann, A. Neurophysiological investigation of the basis of the fMRI signal. *Nature* **412**, 150–157 (2001).
23. Wacongne, C., Changeux, J.-P. & Dehaene, S. A Neuronal Model of Predictive Coding Accounting for the Mismatch Negativity. *J. Neurosci* **32**, 3665–3678 (2012).
24. Jiang, L. P. & Rao, R. P. Dynamic predictive coding: A model of hierarchical sequence learning and prediction in the neocortex. *PLOS Computational Biology* **20**, e1011801 (2024).
25. Friston, K. & Kiebel, S. Predictive coding under the free-energy principle. *Philosophical transactions of the Royal Society B: Biological sciences* **364**, 1211–1221 (2009).
26. Solomon, S. S., Tang, H., Sussman, E. & Kohn, A. Limited evidence for sensory prediction error responses in visual cortex of macaques and humans. *Cereb. Cortex* **31**, 3136–3152 (2021).
27. Yuste, R., Hamm, J., Shymkiv, Y., Han, S. & Yang, W. Cortical ensembles selective for context. *Proceedings of the National Academy of Sciences of USA* **118**, (2021).
28. Hayden, D. J., Montgomery, D. P., Cooke, S. F. & Bear, M. F. Visual recognition is heralded by shifts in local field potential oscillations and inhibitory networks in primary visual cortex. *Journal of Neuroscience* **41**, 6257–6272 (2021).
29. Chen, G. *et al.* Distinct inhibitory circuits orchestrate cortical beta and gamma band oscillations. *Neuron* **96**, 1403–1418 (2017).
30. Gong, Y. *et al.* Neural correlates of novelty detection in the primary auditory cortex of behaving monkeys. *Cell reports* **43**, (2024).
31. Bellet, M. E. *et al.* Spontaneously emerging internal models of visual sequences combine abstract and event-specific information in the prefrontal cortex. *Cell Rep* **43**, (2024).
32. Sotomayor-Gómez, B., Battaglia, F. P. & Vinck, M. Firing rates in visual cortex show representational drift, while temporal spike sequences remain stable. *Cell Rep* **44**, (2025).
33. Maris, E. & Oostenveld, R. Nonparametric statistical testing of EEG- and MEG-data. *J. Neurosci. Methods* **164**, 177–190 (2007).
34. Gavornik, J. P. & Bear, M. F. Learned spatiotemporal sequence recognition and prediction in primary visual cortex. *Nat. Neurosci* **17**, 732–737 (2014).
35. Felleman, D. J. & Essen, D. C. Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* **1**, 1–47 (1991).
36. Lamme, V. A. & Roelfsema, P. R. The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci* **23**, 571–579 (2000).
37. Dhamala, M., Rangarajan, G. & Ding, M. Estimating Granger Causality from Fourier and Wavelet Transforms of Time Series Data. *Phys. Rev. Lett* **100**, 018701 (2008).
38. Nedungadi, A. G., Rangarajan, G., Jain, N. & Ding, M. Analyzing multiple spike trains with nonparametric granger causality. *J. Comput. Neurosci* **27**, 55–64 (2009).
39. Dimidschstein, J. *et al.* A viral strategy for targeting and manipulating interneurons across vertebrate species. *Nat. Neurosci* **19**, 1743–1749 (2016).
40. Torres-Gomez, S. *et al.* Changes in the Proportion of Inhibitory Interneuron Types from Sensory to Executive Areas of the Primate Neocortex: Implications for the Origins of Working Memory Representations. *Cereb. Cortex* **30**, 4544–4562 (2020).
41. Friston, K. A theory of cortical responses. *Philos. Trans. R. Soc. B Biol. Sci* **360**, 815–836 (2005).
42. Qian, N. & Zhang, J. Neuronal Firing Rate As Code Length: a Hypothesis. *Comput. Brain Behav* **3**, 34–53 (2020).
43. Rodriguez, N. Y., McKim, T. H., Basu, D., Ahuja, A. & Desrochers, T. M. Monkey dorsolateral prefrontal cortex represents abstract visual sequences during a no-report task. *Journal of Neuroscience* **43**, 2741–2755 (2023).
44. Murray, J. D. *et al.* A hierarchy of intrinsic timescales across primate cortex. *Nature neuroscience* **17**, 1661–1663 (2014).
45. Wiesel, T. N. & Hubel, D. H. Ordered arrangement of orientation columns in monkeys lacking visual experience. *Journal of comparative neurology* **158**, 307–318 (1974).
46. Chaigneau, E. *et al.* The relationship between blood flow and neuronal activity in the rodent olfactory bulb. *Journal of Neuroscience* **27**, 6452–6460 (2007).
47. Bastos, A. M., Lundqvist, M., Waite, A. S., Kopell, N. & Miller, E. K. Layer and rhythm specificity for predictive routing. *Proc. Natl. Acad. Sci. U. S. A* **117**, 31459–31469 (2020).
48. Nejat, H., Sherfey, J. & Bastos, A. M. Predictive routing emerges from self-supervised stochastic neural plasticity. *bioRxiv* 2024–12 (2025).
49. Findling, C. *et al.* Brain-wide representations of prior information in mouse decision-making. *Nature* **645**, 192–200 (2025).

Methods

Mouse Experiments at the Allen Institute: Animals

All experiments were performed in 16 SSTAi32 and PVAi32 mice (*Mus musculus*) of both sexes (n_{male}=6, n_{female}=10), aged 95-128 days. Optogenetics experiments were conducted with Sst-IRES-Cre/wt; Ai32(RCL-ChR2(H134R)_EYFP)/wt mice (n=9) and Pvalb-IRES-Cre/wt; Ai32(RCL-ChR2(H134R)_EYFP)/wt) mice (n=7). Pvalb-IRES-Cre and Sst-IRES-Cre mice were bred in-house and crossed with an Ai32 channelrhodopsin reporter line. Pvalb-IRES-Cre;Ai32 breeding sets (pairs and trios) consisted of heterozygous Pvalb-IRES-Cre mice crossed with either heterozygous or homozygous Ai32(RCL-ChR2(H134R)_EYFP) mice. Pvalb-IRES-Cre is expressed in the male germline. To avoid germline deletion of the stop codon in the loxP-STOP-loxP cassette, Pvalb-IRES-Cre;Ai32 mice were not used as breeders. Sst-IRES-Cre;Ai32 breeding sets (pairs and trios) consisted of heterozygous Sst-IRES-Cre mice crossed with either heterozygous or homozygous Ai32(RCL-ChR2(H134R)_EYFP) mice. Cre+ cells from Ai32 lines are highly photosensitive, owing to the expression of Channelrhodopsin-2. All experiments on animals were conducted with the approval of the Allen Institute's Institutional Animal Care and Use Committee.

Macaque Monkey Experiments at Vanderbilt University: Animals

All procedures were approved by the Vanderbilt University Institutional Animal Care and Use Committee in compliance with the regulations set by the Association for the Assessment and Accreditation of Laboratory Animal Care and followed the United States National Institutes of Health guidelines. Two macaque monkeys were used in this study: one male adult bonnet macaque (*Macaca radiata*) aged 19 years (Ca, 7.4kg), and one male adult rhesus macaque (*Macaca mulatta*) aged 11 years (Jo, 12.0 kg).

Monkey Ca was injected with AAV vectors carrying an inhibitory cell-specific Dlx-ChR2-mCherry transgene. We mapped the planned trajectory of these injections with co-registered CT and MRI images. While the monkey was awake and head-fixed, we advanced an injectrode (32-channel Plexon V Probe customized with a fluid channel, Plexon, Dallas, TX) connected to a Hamilton syringe (Hamilton Company, Reno, NV) containing the viral vector in areas MT, MST, 8A, and PFC. Each injection site received 3 injections spaced 1mm apart in depth, starting from the deepest location. Injections are followed by a one-minute waiting period, after which the injectrode was retracted to the next injection depth. We injected in total 52.8µL of AAV1-Dlx-ChR2-mCherry and 13.2µL of AAV9-Dlx-ChR2-mCherry (9.6µL AAV1 and 2.4µL AAV9 in PFC, 19.2µL AAV1 and 4.8µL AAV9 in 8a, 14.4µL AAV1 and 3.6µL AAV9 in MT, 14.4µL AAV1 and 3.6µL AAV9 in MST). Each injection was confirmed by simultaneous recording of neuronal activity on nearby channels close to the fluid port, in which neuronal activity decreased during injections.

Mouse Experiments at the Allen Institute: Surgery

The surgery pipeline at the Allen Institute is described in detail in previous reports[12,50]. Briefly, A pre-operative injection of dexamethasone (3.2 mg kg^{-1}, subcutaneously was administered 1 hr before surgery to reduce swelling and postoperative pain by decreasing inflammation. Mice were initially anesthetized with 5% isoflurane (1–3 min) and placed in a stereotaxic frame (Model 1900, Kopf). Isoflurane levels were maintained at 1.5–2.5% for the duration of the surgery. Body temperature was maintained at 37.5 °C. Carprofen was administered for pain management (5–10 mg/kg, subcutaneous), and atropine was administered to suppress bronchial secretions and regulate heart rhythm (0.02–0.05 mg/kg, subcutaneous). An incision was made to remove the skin, and the exposed skull was leveled with respect to pitch (bregma–lambda level), roll, and yaw. The headframe was placed on the skull and fixed with White C&B Metabond (Parkell). Once the Metabond was dry, the mouse was placed in a custom clamp to position the skull at a rotated angle of 20° to facilitate the creation of the craniotomy over the visual cortex. A circular piece of skull 5 mm in diameter was removed, and a durotomy was performed. The brain was covered by a 5-mm-diameter circular glass coverslip, with a 1-mm lip extending over the intact skull. The bottom of the coverslip was coated with a layer of polydimethylsiloxane (SYLGARD 184, Sigma-Aldrich) to reduce adhesion to the brain surface. The coverslip was secured to the skull with Vetbond (Patterson Veterinary). Kwik-Cast (World Precision Instruments) was added around the coverslip to further seal the implant, and Metabond bridges between the coverslip and the headframe well were created to hold the Kwik-Cast in place. At the end of the procedure, but before recovery from anesthesia, the mouse was transferred to a photodocumentation station to capture a spatially registered image of the cranial window.

Intrinsic signal imaging was performed to identify visual area boundaries[51]. An insertion window was designed for each mouse based on the identified visual areas. Mice were habituated to head fixation and visual stimulation over two weeks.

Macaque monkey experiments at Vanderbilt University: Surgery

Monkeys were implanted with a headpost and recording chambers during an anesthetized procedure. Monkeys were anesthetized with isoflurane. Body temperature was maintained near 38°C with a circulating water or warm air recirculating heating blanket. Peripheral oxygen levels, heart rate, temperature, ECG, and expired CO2 levels were monitored continuously during surgery. After mounting the animal in a stereotaxic device, the animal's head was shaved and scrubbed (with Betadine and ethanol). Local anesthetic (Lidocaine) was injected subcutaneously before opening and also placed into the ear canals before fitting the animal into the stereotaxic device.

A skin flap was made over the skull. The skull was cleaned and dried, and orthopedic screws were placed in the skull. The head restraint device, held in position with a stereotaxic holder, was fixed in place with orthopedic screws and methyl methacrylate cement. Custom-made recording chambers (~20 mm in diameter, Crist Instruments) were placed on the skull to allow access to all targeted brain areas in one hemisphere and secured in place with orthopedic screws and dental cement. Craniotomies (10-20 mm diameter) were subsequently performed using piezosurgery drills (Piezosurgery, Columbus, Ohio) to permit access to the underlying cortex. Monkey Ca had two recording chambers, one placed over the prefrontal cortex (giving access to areas FEF and LPFC) and one placed over the temporal cortex (giving access to areas MT, MST, V4, and TEO). Monkey Jo had one recording chamber positioned over the occipital cortex (giving access to areas V1, V2, and V3). After surgery, analgesics were given for pain management (Buprenorphine and NSAIDs for 3 days), and animals showed no signs of pain or distress.

Mouse Experiments at the Allen Institute: Electrophysiology Experiments

The Neuropixels data was acquired at the Allen Institute as part of the OpenScope project that allows the community to apply for observation on the Allen Brain Observatory platform (https://alleninstitute.org/division/mindscope/openscope/). The Neuropixels pipeline at the Allen Institute is described in detail in previous reports[12,50].

On the day of recording, the cranial coverslip was removed and replaced with the insertion window containing holes aligned to six cortical visual areas. Mice were allowed to recover for 1–2 hours after the window was placed before being head-fixed in the recording rig.

Six Neuropixels probes were targeted to each of the six visual cortical areas (V1, LM, RL, AL, PM, AM). The boundaries of these areas and their retinotopic maps were obtained through intrinsic signal imaging, and insertion locations were planned to target regions responsive to the center of the LCD monitor. Probes were doused with CM-DiI (1 mM in ethanol; Thermo Fisher, V22888) for post hoc ex vivo probe localization. Each probe was mounted on a 3-axis micromanipulator (New Scale Technologies). The tip of each probe was aligned to its associated opening in the insertion window using a coordinate transformation obtained via a previous calibration procedure. XYZ manipulator coordinates were obtained via a previous calibration procedure where the tips of probes were aligned to the retinotopic centers of target visual areas using an image provided by intrinsic signal imaging. The operator then moved each probe into place with a joystick, with the probes fully retracted along the insertion axis, approximately 2.5 mm above the brain surface. The probes were manually lowered to the brain surface until spikes were visible on the electrodes closest to the tip. After the probes penetrated the brain to a depth of around 100 µm, they were inserted automatically at a rate of 200 µm/min (total of 3.5 mm or less in the brain). After the probes reached their targets, they were allowed to settle for 5–10 min.

Neuropixels data was acquired at 30 kHz (spike band, 500 Hz high-pass filter) and 2.5 kHz (LFP band, 1,000 Hz low-pass filter) using the Open Ephys GUI[52]. Videos of the eye and body were acquired at 60 Hz. Pupil size was measured via previously described measures[12]. Briefly, a universal eye tracking model trained in DeepLabCut[53], a ResNET-50-based network, recognizes up to 12 tracking points each around the perimeter of the eye, the pupil, and the corneal reflection. The angular velocity of the running wheel was recorded at the time of each stimulus frame, at approximately 60 Hz.

The spike-band data was median-subtracted, first within-channel to center the signal around zero, then across channels to remove common-mode noise. The median-subtracted data file was sent through the Kilosort2 MATLAB package (https://github.com/mouseland/Kilosort[54], which applies a 150 Hz high-pass filter, followed by whitening in blocks of 32 channels. Kilosort2 models this filtered, whitened data as a sum of spike 'templates'. The shape and locations of each template were iteratively refined until the data could be accurately reconstructed from a set of N templates at M spike times. Finally, Kilosort2 output was curated to remove putative double-counted spikes and units with artefactual waveforms. All units not classified as noise were packaged into Neurodata Without Borders (NWB) files[55] and analyzed in this paper.

PV and SOM cells were identified in their respective mouse lines through a brief opto-tagging session at the end of the recordings[56]. Cells were stimulated with an optical fiber at 40 Hz, 5 Hz, and with a raised cosine function in 1 s epochs. Blue light was delivered by a 473 nm laser (Laser Quantum, Stockport, United Kingdom) with a measured light power at the tip of the fiber between 12.5-30 mW. The light source was coupled to a 400 µm diameter fiber optic cable (Thorlabs, Newton, NJ), with the tip positioned such that blue light illuminated the entire cranial window. Light was provided using

three different types of stimuli, randomly interleaved: 2 ms pulses at 5 Hz, 40 Hz, or a 1s raised cosine ramp. Mice had no exposure to the optotagging stimulus prior to the end of the recording session. Cells whose mean response during the stimulation epoch positively exceeded 2 standard deviations of their baseline prestimulus (100 ms epoch) activity to any of the stimulation types were deemed to belong to the associated interneuron group.

Current source density (CSD[57,58] was estimated from the LFP band data by epoching the LFPs, median-filtering with a filter-width of 3 channels, and then downsampling to every 4th channel to retrieve a single vertical/laminar column of responses. CSD was then estimated as the 2nd spatial derivative across multiple contact points according to the equation[59]:

$$CSD(t,c) = -\sigma\,(V(t,c-2) - 2V(t,c) + V(t,c+2))/(2s\hat{}2)$$

V(t, c) is the extracellular voltage at time *t* measured at an electrode contact at position *c*, *s* is the inter-contact distance of the electrode array, and σ is the tissue conductivity (here taken as 0.4 S/m). CSD contrasts were computed by grand-averaging the raw LFPs over each condition, then calculating the CSD as above and subtracting across conditions. We corrected for time in task as a confounder via double contrasts**,** first subtracting the 3rd (fully adapted) presentation response from that of the 4th, and then pooling trials within task conditions. These subtractions were treated as normal random variates for a two-tailed, nonparametric spatiotemporal cluster-test that controls the family-wise error rate to $p < 0.05$ to determine significant cross-condition contrasts[33]. LFP and CSD are smoothed with a width 4 Gaussian filter in both the spatial (laminar) and temporal dimensions for plotting.

Macaque Monkey Experiments at Vanderbilt University: Electrophysiology Experiments

We acquired data using between 1 and 4 linear array multi-channel recording probes. In most experiments, we recorded with 128-channel "deep array" probes where the inter-contact spacing was 25 or 40 µm (Diagnostic Biochips, Glen Burnie, MD) and the coverage was 3.175 mm and 5.08 mm, respectively. In some experiments, we used 32-channel V probes with an inter-contact spacing of 100 µm (Plexon, Dallas, TX) with a coverage of 3.1 mm. We used a mechanical Microdrive (Narishige, Tokyo, Japan) mounted onto the chambers to insert electrodes into the appropriate depth and area. A guide tube was lowered through a recording grid (Crist Instrument Co., Inc.) to penetrate the dura mater and granulation tissue. We then acutely introduced the recording probe into the respective area and waited ~1 hour for the signal to settle before starting the recordings. Recordings were acquired using an RHD System (Intan, Los Angeles, CA) sampling at 30 kHz. Recordings were referenced to the guide tube, which remained in contact with tissue. Grid positions were pre-determined based on a pre-recording MRI scan with the recording chambers and grid in place, with water to mark the trajectories of each grid hole position. We advanced the electrode until there was visually responsive neuronal activity on most channels and until the local field potentials showed a distinct spectral signature of the cortical sheet, characterized by large amplitude alpha/beta (10-30 Hz) oscillations in deep channels and gamma (40-150 Hz) oscillations in superficial channels[60–64]. Offline, using MATLAB, the data was filtered between 500-5000 Hz, full-wave rectified, low-pass filtered below 250 Hz, and downsampled to 1000 Hz. We refer to this signal as the MUAe signal, and it reflects the activity of nearby surrounding neurons[22,62,65–68].

We allowed three to seven weeks after the optogenetic injections for viral expression before stimulating and recording. Monkey Ca was awake and head-fixed during opto-stimulation. A DBC 128-channel probe with a fiber (an "optrode") attachment (0.2mm in diameter with a numerical aperture of 0.37) at the end was connected to a laser with a wavelength of 473 nm and output power of 318 mW (before fiber). Laser power at the fiber tip was tested in a calibration run with a photodiode at 92.1 mW maximum (SM1-Threaded Mounted Photodiode, Thorlab, Newton, New Jersey) in an Optical Power Monitor (Thorlab, Newton, New Jersey). On recording days, the optrode was advanced into the injection site (in MT, MST, 8a, and PFC in separate sessions) with a metal guide tube puncturing the dura mater. Electrophysiological signals were recorded on the same optrode and a non-optrode neural probe simultaneously in another area. For analysis of the monkey optogenetics dataset, we used isolated single units and used the same spike sorting techniques that were applied to mouse data (Kilosort2).

Neurons were stimulated with a 5 Hz and 40 Hz raised cosine function at 3V, 4V, 5V amplitudes, at maximum output power of 13.9 mW, 71.2 mW, and 92.1 mW, respectively. Instantaneous firing rate was estimated by smoothing binary spike trains, using a nonparametric EPSP filter (20ms smoothing window). Mean spike waveforms, attributes (trough-peak duration, voltage distribution along the probe), and quality metrics (presence ratio, overall firing rate, signal-to-noise-ratio, and inter-spike interval violations) were manually checked for each Kilosort2 sorted optotagged unit, and noisy/artifactual units based on these attributes were removed prior to analysis. Each single unit's firing rate during laser stimulation (2000 ms) was compared with baseline (1000 ms before onset and 1000 ms after offset). Trial-averaged mean firing rates for each laser frequency and power level were compared with baseline using one-tailed Wilcoxon signed-rank tests ($\alpha = 0.05$, right-tailed). Magnitude-squared coherence between the laser and each unit was computed across the spectrum of all laser stimulation

frequencies. Coherence was defined as significant under two criteria: if coherence of spiking and laser during laser interval was higher than for spiking during the pre- and post-baseline, and if coherence surpassed a critical values[69]. Units significant for increased firing rate and coherence were considered optotagged.

Mouse Experiments at the Allen Institute: Visual Stimulation Parameters

All Neuropixels recordings were according to the standardized Neuropixels visual coding pipeline. Visual stimuli were generated using custom PsychoPy scripts and displayed using an ASUS PA248Q LCD monitor with 1920 × 1200 pixels (55.7 cm wide, 60 Hz refresh rate). Stimuli were presented monocularly, and the monitor was positioned 15 cm from the right eye of the mouse and spanned 120 dva × 95 dva. Monitor placement was standardized across rigs such that the mouse's right eye gaze would typically be directed at the center of the monitor. Then, a sequence of drifting gratings appeared (full-screen, 4 Hz drift rate, 1 cycle/degree, 0.8 Michelson contrast, angle of the drifting grating was 45 or 135 degrees from the vertical midline) for 500 ms, and was replaced by a blank screen with the fixation dot remaining for 500 ms. This sequence was repeated 4 times in a trial, for a total trial length of 4500 ms (500 ms pre-sequence fixation + 4 x 1000 ms stimulus presentations). Every 100 sequences, a period of a blank, black screen was presented for 10 s.

Macaque Monkey Experiments at Vanderbilt University: Visual Stimulation Parameters

We used MonkeyLogic (developed and maintained at the National Institute for Mental Health) to control the behavioral task[70]. Visual stimuli were displayed using PROPixx Pro projectors (VPixx Technologies, Quebec, Canada) with a resolution of 1920 x 1080 at a 120 Hz refresh rate. The projector screen was positioned 57 cm in front of the monkey's eyes. Luminance calibration was performed with a Photo Research PR-650 at the projector screen with a median value of 2.613 cd/m^2.

The monkey was trained to fixate its eyes around a central fixation dot (radius of fixation window: 1.5 dva). Eye position and pupil diameter were monitored using an Eyelink 1000 (SR Research, Ottawa, Canada) at 250 or 500 Hz. A task began with an isoluminant gray screen. Once the monkey maintained fixation for 500ms, a sequence of drifting gratings appeared (12 visual degree radius, 2 Hz drift rate, 1 cycle/degree, 0.8 Michelson contrast, angle of the drifting grating was 45 or 135 degrees from the vertical midline) for 500 ms, and was replaced by a blank screen with the fixation dot remaining for 500 ms. This sequence was repeated 4 times in a trial, for a total trial length of 4500 ms (500 ms pre-sequence fixation + 4 x 1000 ms stimulus presentations). The monkey was rewarded with a small juice reward for maintaining fixation throughout the trial.

Mouse Experiments at the Allen Institute: Habituation to Predicted Sequence

Mice were habituated over the course of 5 behavior-only sessions in which only the predicted sequence (either x-x-x-y or y-y-y-x) was presented. This resulted in 2000 habituation trials before recordings.

Macaque Monkey Experiments at Vanderbilt University: Habituation to Predicted Sequence

Monkeys were habituated to a predicted sequence over the course of 3 behavior-only sessions in which only the predicted sequence (either x-x-x-y or y-y-y-X) was presented. This resulted in 2850 complete habituation trials before recording sessions. Subsequently, after a recording session, monkeys were again re-habituated to one of the predicted sequences before the next recording session.

Mouse Experiments at the Allen Institute: Data Analysis

We recorded 29977 total putative single units, averaging 359 (+/-42) units per animal and 60 (+/-3) units per probe. We further restricted the units for analysis, consistent with previous reports using the same methods for analysis[12], to those that exhibit a visual response that exceeded 5 standard deviations of the baseline firing rate, minimum of 0.1 sp/s, inter-spike interval (ISI) violations below 0.33, a presence ratio above 0.95, an amplitude cutoff below 0.1, and were found to be in areas V1, LM, RL, AL, PM, AM, or LP. This resulted in a sample of 5389 units from 15 of 16 mice. No units from 1 mouse passed the criteria due to a destabilizing event during the recording. Additionally, data from another mouse were omitted from the analysis due to issues with the visual stimulus timing. As a result, our final sample consisted of 5062 single units from 14 mice.

Macaque Monkey Experiments at Vanderbilt University: Data Analysis

We recorded a total of 6,619 sites with MUAe signals across 8 cortical areas across 19 recording sessions: 640 in PFC, 416 in FEF, 512 in MST, 512 in MT, 1056 in V4, 1536 in V3, 896 in V2, and 1024 in V1. These MUAe signals were extracted from broadband signals with a Butterworth filter of stopbands at 450 Hz and 5500 Hz and passbands at 500 Hz and 5000 Hz. For analysis, we restricted these sites that had visually evoked activity, which was determined by whether the period from 50-250 ms post-stimulus was significantly greater than pre-stimulus baseline (using a threshold of $p<0.05$, rank sum

test). The resulting population used for analysis consisted of 508 sites in PFC, 177 in FEF, 161 in MST, 252 in MT, 230 in V4, 569 in V3, 328 in V2, and 270 in V1.

Macaque Monkey Experiments at Vanderbilt University: Laminar Identification

For each penetration of monkey recordings, we examined its trajectory based on the MRI image and postmortem Nissl staining images. Penetrations that are not sufficiently perpendicular to cortical layers or did not have enough cortical coverage were excluded from layer-based analyses. We then examined the multi-unit activity, CSD, and spectrolaminar pattern[61] across depth for each penetration. Layer 4 of each area was identified based on the best corresponding electrode location with the earliest MUA onset, earliest CSD sink, and spectrolaminar high/low band relative power crossing. Total area thickness was measured in MRI images of each monkey for each area. We also obtained percent thickness per layer for each area from existing data[71]. The number of electrodes per layer was then calculated by multiplying the total layer thickness by the thickness of each layer for each area. We then assigned the layer position of each electrode by counting from the previously identified layer 4 electrode up to layer 1 and down to layer 6.

Granger causality analysis

In mice, we began by defining all available single units in each area. We took all cortical neurons for this analysis, and per mouse and probe, included an area in the analysis if a minimum of 12 single neurons were simultaneously acquired. This resulted in six neuronal spiking time series, one per area. Granger causality was calculated separately per mouse and for local and global oddball trials, using all available trials. We calculated Granger causality using a non-parametric implementation[37,38]. For this, we computed Fourier coefficients using the Fieldtrip toolbox[72] using a sliding window approach, utilizing 200 ms time windows for analysis. We computed the Fourier coefficients using multitapers (3 multitapers per trial, time window size = 200 ms, smoothing = 10 Hz, function 'ft_freqanalysis', using method 'mtmfft', frequencies 0-100 Hz were calculated). One time window was on the baseline, from 100 ms pre- and 100 ms post-oddball onset. Two time windows reflecting stimulus processing were also used, one centered from 100 ms to 300 ms post-oddball onset, another from 300 ms to 500 ms post-oddball onset. Bivariate Granger causality was computed pair-wise between all available neuronal spiking time series (using ft_connectivity analysis with method = 'granger' and granger.sfmethod = 'bivariate').

Statistics and Visualization

In mice, spiking activity for each cell was normalized by their average activity level to produce an arbitrary value that was less variable across units with differing firing rates. In monkeys, spiking activity for each channel was normalized to each channel's standard error across trials. Unless otherwise noted, all statistics were performed using nonparametric, cluster-based permutation testing[33] of the spiking activity at either the single-(multi)unit level or the population level, with alpha thresholds set to 0.05 for both test stages. We also required that a statistically significant difference persist for at least 25 ms as indicated by this test. Data was smoothed with a 50-ms moving mean window. Confidence bands, where displayed, are computed for a 95% confidence interval across units' condition-wise mean activity included in the population.

Feature selectivity was measured the same way as in a previous reports[73]. First, we determined the mean response to each stimulus during visual stimulation (0-500 ms following stimulus appearance) and during a separate presentation block where stimulus order was randomized. For each unit, we subtracted the mean response to the GO-orientation stimulus from the LO-orientation stimulus and divided it by the sum of those two means. Single units with this index above 0.8 or below -0.8 were defined as highly selective for the GO orientation or the LO orientation, respectively.

For identifying the onset latencies of significance, a t-test was run at each time point during 500 ms of stimulus presentation for individual unit/channel (with the same selection criteria described above, only including unit/channel with significant visually evoked activity), and the first contiguous significant cluster larger than 10ms was considered the onset of significance for that channel.

Data and Code Availability

All mouse data used in this work are openly available through the DANDI Archive (dandiset id: 000253, doi: https://doi.org/10.48324/dandi.000253/0.240503.0152). Software resources for the analysis of this dataset are provided here: https://alleninstitute.github.io/openscope_databook/projects/glo.html. The code to generate the global-local oddball sequences is provided here: https://github.com/AllenInstitute/openscope-glo-stim.