Meta-reinforcement learning in a thalamo-orbitofrontal circuit

Vijay Mohan K Namboodiri[1], Taylor Hobbs[1], Ivan Trujillo Pisanty[1], Rhiana C Simon[1], Garret D Stuber[1, 2]

[1] The Center for the Neurobiology of Addiction, Pain, and Emotion, Department of Anesthesiology and Pain Medicine, Department of Pharmacology, The Center for the Neurobiology of Addiction, Pain, and Emotion, University of Washington, Seattle, WA 98199, USA

[2] Corresponding author: Email: gstuber@uw.edu

Learning to predict rewards is essential for the survival of animals. Contemporary views suggest that such learning is driven by a reward prediction error—the difference between received and predicted rewards. Here we show using two-photon calcium imaging and optogenetics in mice that a different class of reward learning signals exists within the orbitofrontal cortex (OFC). Specifically, the reward responses of many OFC neurons exhibit plasticity consistent with filtering out rewards that are less salient for learning (such as predicted rewards, or, unpredicted rewards available in a context containing highly salient aversive stimuli). We show using quasi-simultaneous imaging and optogenetics that this reward response plasticity is sculpted by medial thalamic inputs to OFC. These results provide a biological substrate for emerging theoretical views of meta-reinforcement learning in prefrontal cortex.

A simple, yet powerful model for learning that a cue predicts an upcoming reward is to update one's predictions of reward by a reward prediction error (RPE)—the difference between a received and predicted reward (Rescorla and Wagner, 1972; Schultz et al., 1997). This RPE model has been successful at explaining the response dynamics of midbrain dopaminergic neurons (Eshel et al., 2015; Mohebi et al., 2019; Schultz et al., 1997)—the main known neural system controlling reward prediction learning (Chang et al., 2016; Steinberg et al., 2013). Even though the bulk of experimental work into the neuronal mechanisms of learning focus on the mesolimbic dopamine circuitry, a rich computational literature argues that the dopamine RPE circuitry cannot operate in isolation and that the brain likely contains complementary systems for learning (Schweighofer and Doya, 2003; Soltani and Izquierdo, 2019; Wang et al., 2018). One such theoretical hypothesis is that of "meta-reinforcement learning"—the learning of parameters needed for learning (Iigaya, 2016; Schweighofer and Doya, 2003; Wang et al., 2018). For instance, prior computational models have proposed that parameters such as the learning rate (the magnitude by which a prediction error affects prediction) are themselves learned (Behrens et al., 2007; Iigaya, 2016; Schweighofer and Doya, 2003; Soltani and Izquierdo, 2019; Wang et al., 2018). Such a meta-reinforcement learning system, hypothesized to exist in the prefrontal cortex (Soltani and Izquierdo, 2019; Wang et al., 2018), should be capable of identifying when a reward should have a low learning rate. In this paper, we show using longitudinal activity tracking by two-photon calcium imaging that some neuronal subpopulations in the ventral/medial orbitofrontal cortex (vmOFC) reverse the sign of their reward responses from positive to negative to reflect when rewards should have low learning rates. We also establish that this reward response plasticity is controlled by medial thalamic input to vmOFC. We thus demonstrate a biological substrate of meta-reinforcement learning at single neuron resolution in a thalamo-prefrontal cortical network. Our results bolster the emerging theoretical view that the prefrontal cortex acts as a meta-reinforcement learning system (Wang et al., 2018), in addition to representing cognitive parameters for learning such as uncertainty (Soltani and Izquierdo, 2019), confidence (Kepecs et al., 2008) and volatility (Behrens et al., 2007).

Prior theoretical meta-reinforcement learning models have proposed that the learning rate of a reward should be tuned down as the reward becomes predicted in a static environment (Iigaya, 2016;

Schweighofer and Doya, 2003). To investigate reward response plasticity in a large number of individual vmOFC neurons during reward prediction learning, we used two-photon calcium imaging during a discriminative Pavlovian trace conditioning task in head-fixed mice (Namboodiri et al., 2019; Otis et al., 2017) (**Fig 1A, B**). This allowed us to test if reward receipt responses change during learning in a manner consistent with signaling that predicted rewards are less salient for learning. We previously showed that vmOFC neurons cluster into different neuronal subpopulations based on their time-locked activity profiles to the CS+ (auditory cue paired with delayed reward) and CS- (auditory cue not paired with reward) (**Fig S1**) (Namboodiri et al., 2019). Early in learning the CS+-reward association, the response to reward was positive in all clusters, including in OFC neurons projecting to VTA (OFC→VTA) (**Fig S1**). However, once mice learned to predict the upcoming reward upon sensing CS+, there was a reduction in the reward responses of most clusters (**Fig 1C**, **Fig S1**). Indeed, some clusters (1 and 3) displayed a negative reward response late in learning, even though the same neurons showed positive reward responses early in learning (**Fig 1C**, **Fig S1**). Importantly, while the reward response early in learning only reflects reward receipt, reward response late in learning is the sum of both a reward receipt and a reward prediction response (**Fig 1D**, see next paragraph for how this compares to RPE). Thus, to isolate the reward receipt response late in learning, we reduced the probability of reward associated with CS+ from 100% to 50%. In this session, the pure reward receipt response is the difference in response between the rewarded and unrewarded trials (**Fig 1E**). We found using this approach that the immediate reward receipt response (0-3 s after reward) late in learning was negative for clusters 1, 3 and 6, resulting in a flip in the sign of the response from early in learning (**Fig 1F, G**). Though clusters 2 and 5 showed significant positive reward receipt response in the 50% probability session, the response magnitude for cluster 2 was much smaller than the response early in learning (**Fig 1 G**, **Fig S1**). These results show that vmOFC reward responses are highly plastic during reward prediction learning, with several distinct subpopulations (clusters 1, 2, 3 and 6) showing large reduction across learning.

Reward responses are also plastic in midbrain dopaminergic neurons, whose large and positive reward responses early in learning become significantly lower later in learning (Coddington and Dudman, 2018; Hollerman and Schultz, 1998). This is commonly thought to reflect a linear subtraction of reward receipt and reward prediction (i.e. RPE) (Eshel et al., 2015). Linear models such as RPE cannot explain our observations in vmOFC, as the pure reward receipt response (reward response *minus reward prediction response*) in these models do not change across learning. However, in the above data, the pure reward receipt response changes significantly due to reward prediction (**Fig 1G**). The simplest model consistent with these data is a nonlinear extension in which reward receipt response is itself negatively gated by reward prediction (**Fig S1**). Prior computational models have proposed that parameters such as the learning rate (amount by which prediction error changes prediction) are themselves learned (i.e. “meta-learning”) and are tuned down over the course of learning in a static environment (Iigaya, 2016; Schweighofer and Doya, 2003; You et al., 2019). The nonlinear gating model mentioned above operates similarly by tuning down reward receipt responses by reward prediction. Consistent with this view of vmOFC reward responses setting learning rates, we previously found that suppressing the reward, but not cue, responses of OFC→VTA neurons, reduces behavioral learning rate based on recent reward history (Namboodiri et al., 2019). Overall, these results are consistent with a meta-reinforcement learning signal in vmOFC that reduces the learning rate of predicted rewards.

We next tested if vmOFC reward responses are consistent with meta-reinforcement learning in a setting independent of reward prediction. Prior meta-reinforcement learning models have proposed that learning rates are dynamically and individually set for different inputs (Sutton, 1992). Animals typically have higher learning rates for punishments than for rewards (Frank et al., 2009; Galea et al., 2015; Gershman, 2015; Kojima et al., 1996), suggesting that animals prioritize salient aversive stimuli over rewards for learning. For instance, prediction of highly salient aversive stimuli such as foot shocks or

quinine (a bitter tastant) often occurs in single trials (Ader et al., 1972; Slotnick and Coppola, 2015). We thus hypothesized that delivering rewards in a context that also includes the delivery of salient aversive stimuli would result in a suppression of vmOFC reward responses due to a relative reduction in the salience of sucrose, *independent* of the suppression due to reward prediction (see **Supplementary Note 1**). To minimize sensory confounds, we used an aversive stimulus delivered using the same sensory modality as the reward (i.e. taste). We thus intermittently and randomly (i.e. unpredictable) delivered drops of either sucrose or high concentration quinine (1.5-2.5 mM) in a 3:1 ratio to headfixed mice (**Fig 2A**). Since the liquid deliveries were unpredictable and mostly sucrose, the animals consistently sampled the liquid to ascertain whether a given drop is rewarding or aversive. On a given trial, mice quickly stopped licking if the liquid was quinine, demonstrating aversion (**Fig 2B, C**). This experiment allowed us to test whether sucrose responses change in a context containing quinine. Consistent with our hypothesis, we found that unpredicted sucrose responses produced negative responses in vmOFC neurons when delivered in a context containing quinine (**Fig 2D**). In comparison, reward responses early in learning (**Fig 2D**, **Fig 1D**), and unpredicted reward responses after learning in the absence of quinine (**Fig S2**), were positive across all clusters. We reproduced these results in OFC→VTA neuronal subpopulations (**Fig S2**). Hence, even when reward delivery is unpredicted, vmOFC reward responses are reduced by the presence of unpredicted salient aversive stimuli (see **Supplementary Note 2**). Thus, consistent with a meta-reinforcement learning system that signals when rewards are less salient for learning, vmOFC reward responses reduce in two independent settings due to reward prediction or the reduction of relative salience of unpredicted rewards (see **Supplementary Note 3**).

We next assessed the neuronal circuit mechanism for such meta-reinforcement learning in vmOFC neurons. We hypothesized that reward responses in vmOFC may be at least partially controlled by inputs from medial thalamus (mThal). This is because a wide array of reward responsive regions such as basolateral amygdala, other prefrontal cortical regions, and pallidal regions project to the medial thalamus and can indirectly control vmOFC activity through mThal (Jankowski et al., 2013; Mitchell and Chakraborty, 2013). Further, disconnection studies have shown that interactions between mThal and OFC are necessary for reward related decision-making (Izquierdo and Murray, 2010). Despite this, whether reward responses in mThal→vmOFC input exhibit meta-reinforcement learning and whether this input causally affects meta-reinforcement learning in vmOFC is unknown. We first identified the anatomical locations of thalamic cell bodies projecting to vmOFC using a viral (Tervo et al., 2016) and non-viral (Otis et al., 2017) retrograde tracing approaches (**Fig 3A**). The predominant thalamic structures projecting to vmOFC are the anteromedial and mediodorsal thalamic nuclei (**Fig 3B**, **Fig S3**). We then investigated the reward response plasticity of input from these regions to vmOFC. We compared unpredicted reward responses of mThal→vmOFC axons in sessions without and with quinine to test for plasticity consistent with meta-reinforcement learning (**Fig 3C-E**). We found positive and heterogeneous activity patterns in these axons in response to unpredicted sucrose rewards in the absence of quinine (**Fig 3F**, **Fig S3**). These reward responses were suppressed in a session containing quinine (**Fig 3F**), showing qualitative correspondence with the meta-reinforcement learning observed in vmOFC neuronal responses.

These results suggest that mThal input might contribute to the reward response plasticity observed in vmOFC neurons. To test the causal influence of mThal input on vmOFC reward responses, we optogenetically inhibited mThal→vmOFC input after reward delivery while imaging from vmOFC neurons (**Fig 4A**). To remove light artifacts, we discarded the imaging frames during optogenetic inhibition, and evaluated reward responses right after the termination of inhibition. Since GCaMP6s responses are slow with a decay time of roughly two seconds (Chen et al., 2013), a change in activity during the inhibition will be apparent even after the inhibition for up to two seconds. We found that individual neurons showed both positive and negative modulation of activity due to mThal inhibition (**Fig 4B**). On average, clusters 1, 2, 3 and 6, showed increased reward responses due to mThal

inhibition (**Fig 4C, D**). Interestingly, these were the exact same clusters that showed negative reward response gating by reward prediction (**Fig 1G**), suggesting that the negative gating depends on mThal input. We ruled out confounds due to rebound excitation of axons, or the presence of light as no effect was observed during spontaneous inhibition in the absence of rewards or in virus control animals without opsin expression (**Fig S4**). To directly test whether mThal input affects reward response plasticity, we inhibited mThal axons in a session containing unpredicted deliveries of sucrose and quinine. We found that the reduction in sucrose responses due to the presence of quinine was significantly dampened upon mThal inhibition in all clusters except clusters 4 and 5 (**Fig 4F**). There was also a non-selective change in quinine responses in some clusters (**Fig S4**). Lastly, we found that the reward response plasticity due to reward prediction (**Fig 1**) or the presence of quinine (**Fig 2**), predicts the average effect of mThal inhibition on a given vmOFC neuronal subpopulation (**Fig 4 G, H**). Overall, these results show that the meta-reinforcement learning in vmOFC neurons depends heavily on mThal inputs.

The results presented here demonstrate that vmOFC neurons show large reward response plasticity in a manner consistent with adjusting their activity to signal when rewards should have low learning rates. A simple mathematical model for such plasticity is shown in **Fig 4I**. Meta-reinforcement learning algorithms propose that parameters such as learning rates are themselves learned (Schweighofer and Doya, 2003; Sutton, 1992; Wang et al., 2018), potentially in the prefrontal cortex (Wang et al., 2018). Our results demonstrate a biological substrate supporting meta-reinforcement learning. Consistent with meta-reinforcement learning, we previously found that inhibition of OFC→VTA neurons during the reward, but not the cue, suppresses behavioral learning rate dependent on recent reward history (Namboodiri et al., 2019). A previous study also found that inhibition of lateral OFC (likely also containing some ventral OFC) during the reward, but not cue, period of an instrumental task, affected behavioral adaptation dependent on reward history (Constantinople et al., 2019). Further, lesions of medial OFC affect learning and representation of outcomes associated with an action (Bradfield and Hart, 2020). This finding is potentially consistent with meta-reinforcement learning, as lesion of medial OFC may cause “over-learning” of an action, thereby making it less sensitive to the expected outcome. While the above results are consistent with a role for vmOFC neurons in setting learning rate, careful future experiments are needed to rule out alternative explanations, and to selectively, but bilaterally, inhibit the subpopulations showing reward response gating. Future experiments are also needed to test whether vmOFC reward responses reflect the salience of rewards for learning in a dynamic and uncertain environment (Behrens et al., 2007; Iigaya, 2016; Soltani and Izquierdo, 2019), thereby functioning more generally as a system that prioritizes currently available rewards or punishments for learning. It is also important to test the generality of these findings to different brain regions and types of reward/punishment. For instance, it is unclear if these results are unique to vmOFC and mThal. Since thalamic outputs are excitatory (Halassa and Sherman, 2019), the results presented here likely imply a key role for vmOFC inhibitory interneurons in shaping the responses of vmOFC output neurons. This means that while mThal inputs are integral for the reward response plasticity in vmOFC output neurons, the computation happens within the vmOFC local circuit. Considering that mThal input is itself heterogeneous (**Fig S3**), teasing out the exact circuit mechanism by which the different streams of information are integrated within the different subpopulations of vmOFC will be highly technically challenging (**Fig 4I**). Additional experiments are also necessary to identify how these vmOFC signals work together with midbrain dopaminergic neurons signaling RPE (Schultz et al., 1997; Steinberg et al., 2013) to control reinforcement learning in the brain. An intriguing possibility is that the reduction of reward responses in vmOFC neurons after learning may counteract the significant dopaminergic reward responses observed even in extensively trained animals (Engelhard et al., 2019; Lee et al., 2020). This balance between a low learning rate signaled by vmOFC and a significantly positive RPE

signaled by dopaminergic neurons, may be especially important for predicting delayed rewards and accounting for uncertainty in subjective time. While our results highlight the need for future experiments, they provide a biological foundation to investigate the neuronal circuit mechanisms of meta-reinforcement learning.

References

Ader, R., Weijnen, J.A.W.M., and Moleman, P. (1972). Retention of a passive avoidance response as a function of the intensity and duration of electric shock. Psychon Sci *26*, 125–128.

Behrens, T.E.J., Woolrich, M.W., Walton, M.E., and Rushworth, M.F.S. (2007). Learning the value of information in an uncertain world. Nat Neurosci *10*, 1214–1221.

Bradfield, L.A., and Hart, G. (2020). Rodent medial and lateral orbitofrontal cortices represent unique components of cognitive maps of task space. Neurosci Biobehav Rev *108*, 287–294.

Chang, C.Y., Esber, G.R., Marrero-Garcia, Y., Yau, H.-J., Bonci, A., and Schoenbaum, G. (2016). Brief optogenetic inhibition of dopamine neurons mimics endogenous negative reward prediction errors. Nat Neurosci *19*, 111–116.

Chen, T.-W., Wardill, T.J., Sun, Y., Pulver, S.R., Renninger, S.L., Baohan, A., Schreiter, E.R., Kerr, R.A., Orger, M.B., Jayaraman, V., et al. (2013). Ultrasensitive fluorescent proteins for imaging neuronal activity. Nature *499*, 295–300.

Coddington, L.T., and Dudman, J.T. (2018). The timing of action determines reward prediction signals in identified midbrain dopamine neurons. Nat. Neurosci. *21*, 1563–1573.

Constantinople, C.M., Piet, A.T., Bibawi, P., Akrami, A., Kopec, C., and Brody, C.D. (2019). Lateral orbitofrontal cortex promotes trial-by-trial learning of risky, but not spatial, biases. Elife *8*.

Engelhard, B., Finkelstein, J., Cox, J., Fleming, W., Jang, H.J., Ornelas, S., Koay, S.A., Thiberge, S.Y., Daw, N.D., Tank, D.W., et al. (2019). Specialized coding of sensory, motor and cognitive variables in VTA dopamine neurons. Nature *570*, 509–513.

Eshel, N., Bukwich, M., Rao, V., Hemmelder, V., Tian, J., and Uchida, N. (2015). Arithmetic and local circuitry underlying dopamine prediction errors. Nature *525*, 243–246.

Frank, M.J., Doll, B.B., Oas-Terpstra, J., and Moreno, F. (2009). Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. Nat Neurosci *12*, 1062–1068.

Galea, J.M., Mallia, E., Rothwell, J., and Diedrichsen, J. (2015). The dissociable effects of punishment and reward on motor learning. Nat. Neurosci. *18*, 597–602.

Gershman, S.J. (2015). Do learning rates adapt to the distribution of rewards? Psychon Bull Rev *22*, 1320–1327.

Halassa, M.M., and Sherman, S.M. (2019). Thalamocortical Circuit Motifs: A General Framework. Neuron *103*, 762–770.

Hollerman, J.R., and Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. Nat. Neurosci. *1*, 304–309.

Iigaya, K. (2016). Adaptive learning and decision-making under uncertainty by metaplastic synapses guided by a surprise detection system. ELife *5*, e18073.

Izquierdo, A., and Murray, E.A. (2010). Functional interaction of medial MD thalamus but not nucleus accumbens with amygdala and orbital prefrontal cortex is essential for adaptive response selection after reinforcer devaluation. J Neurosci *30*, 661–669.

Jankowski, M.M., Ronnqvist, K.C., Tsanov, M., Vann, S.D., Wright, N.F., Erichsen, J.T., Aggleton, J.P., and O'Mara, S.M. (2013). The anterior thalamus provides a subcortical circuit supporting memory and spatial navigation. Front Syst Neurosci *7*.

Kepecs, A., Uchida, N., Zariwala, H.A., and Mainen, Z.F. (2008). Neural correlates, computation and behavioural impact of decision confidence. Nature *455*, 227–231.

Kojima, S., Yamanaka, M., Fujito, Y., and Ito, E. (1996). Differential Neuroethological Effects of Aversive and Appetitive Reinforcing Stimuli on Associative Learning in Lymnaea stagnalis. Jzoo *13*, 803–812.

Lee, K., Claar, L.D., Hachisuka, A., Bakhurin, K.I., Nguyen, J., Trott, J.M., Gill, J.L., and Masmanidis, S.C. (2020). Temporally restricted dopaminergic control of reward-conditioned movements. Nat. Neurosci. *23*, 209–216.

Mitchell, A.S., and Chakraborty, S. (2013). What does the mediodorsal thalamus do? Front Syst Neurosci *7*, 37.

Mohebi, A., Pettibone, J.R., Hamid, A.A., Wong, J.-M.T., Vinson, L.T., Patriarchi, T., Tian, L., Kennedy, R.T., and Berke, J.D. (2019). Dissociable dopamine dynamics for learning and motivation. Nature *570*, 65–70.

Namboodiri, V.M.K., Otis, J.M., Heeswijk, K. van, Voets, E.S., Alghorazi, R.A., Rodriguez-Romaguera, J., Mihalas, S., and Stuber, G.D. (2019). Single-cell activity tracking reveals that orbitofrontal neurons acquire and maintain a long-term memory to guide behavioral adaptation. Nat. Neurosci. *22*, 1110.

Otis, J.M., Namboodiri, V.M.K., Matan, A.M., Voets, E.S., Mohorn, E.P., Kosyk, O., McHenry, J.A., Robinson, J.E., Resendez, S.L., Rossi, M.A., et al. (2017). Prefrontal cortex output circuits guide reward seeking through divergent cue encoding. Nature *543*, 103–107.

Rescorla, R.A., and Wagner, A.R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. Classical Conditioning II: Current Research and Theory *2*, 64–99.

Schultz, W., Dayan, P., and Montague, P.R. (1997). A Neural Substrate of Prediction and Reward. Science *275*, 1593–1599.

Schweighofer, N., and Doya, K. (2003). Meta-learning in reinforcement learning. Neural Netw *16*, 5–9.

Slotnick, B., and Coppola, D.M. (2015). Odor-Cued Taste Avoidance: A Simple and Robust Test of Mouse Olfaction. Chem Senses *40*, 269–278.

Soltani, A., and Izquierdo, A. (2019). Adaptive learning under expected and unexpected uncertainty. Nat Rev Neurosci *20*, 635–644.

Steinberg, E.E., Keiflin, R., Boivin, J.R., Witten, I.B., Deisseroth, K., and Janak, P.H. (2013). A causal link between prediction errors, dopamine neurons and learning. Nat Neurosci *16*, 966–973.

Sutton, R.S. (1992). Adapting Bias by Gradient Descent: An Incremental Version of Delta-Bar-Delta. AAAI 171–176.

Tervo, D.G.R., Hwang, B.-Y., Viswanathan, S., Gaj, T., Lavzin, M., Ritola, K.D., Lindo, S., Michael, S., Kuleshova, E., Ojala, D., et al. (2016). A Designer AAV Variant Permits Efficient Retrograde Access to Projection Neurons. Neuron *92*, 372–382.

Wang, J.X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J.Z., Hassabis, D., and Botvinick, M. (2018). Prefrontal cortex as a meta-reinforcement learning system. Nat. Neurosci. *21*, 860–868.

You, K., Long, M., Wang, J., and Jordan, M.I. (2019). How Does Learning Rate Decay Help Modern Neural Networks? ArXiv:1908.01878 [Cs, Stat].

Acknowledgments

We thank P. Phillips, K. Ishii, R. Gowrishankar, S. Piantadosi, M. Gray, J. Rodriguez-Romaguera and Z.C. Zhou for comments on the manuscript, and S. Mihalas and all Stuber lab members for helpful discussions. We thank Karl Deisseroth (Stanford University) and the GENIE project at Janelia Research Campus for viral constructs. **Funding:** This study was funded by grants from the National Institute of Mental Health (K99MH118422, V.M.K.N.; F31 MH117931, R.S.), National Institute of Drug Abuse (R37-DA032750 & R01-DA038168, G.D.S.) and Brain and Behavior Research Foundation (NARSAD Young Investigator Award, V.M.K.N.). **Author Contributions:** V.M.K.N., T.G.H., I.T.P., and R.S. performed experiments. V.M.K.N. performed analyses. V.M.K.N., and G.D.S. designed experiments and wrote the manuscript with input from all authors. **Competing Interests:** None. **Data and materials availability:** The data used in this study are available from the corresponding author upon request. All of the behavioral data were collected using custom MATLAB and Arduino scripts written by VMKN. These are available upon request from the corresponding author. All of the analysis was done in Python using custom codes written by VMKN. These will be uploaded to the Stuber lab Github page (https://github.com/stuberlab), and/or will be available upon request from the corresponding author.

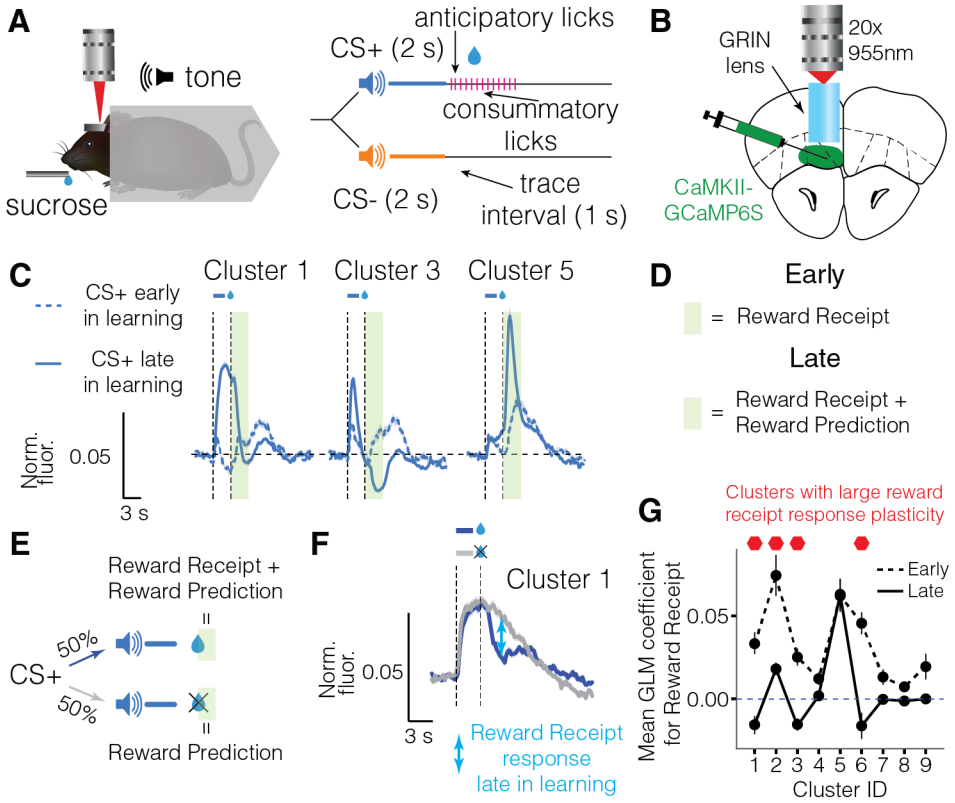


Fig 1: Reward responses in some vmOFC neuronal subpopulations are highly plastic during reward prediction learning.

- **A.** Differential trace conditioning task in headfixed mice (Namboodiri et al., 2019; Otis et al., 2017).
- **B.** Schematic of surgery for two-photon calcium imaging.
- **C.** Peri-Stimulus Time Histograms (PSTHs) early and late in learning for average CS+ responses across all CaMKIIα expressing neurons in clusters 1, 3 and 5 (see **Fig S1** for all clusters). The identification of neuronal subpopulations using clustering algorithms and the PSTHs late in learning were published previously (Namboodiri et al., 2019). The average response includes the responses of all recorded neurons. The green shaded region corresponds roughly to the epoch over which reward responses were calculated, with the actual epoch corresponding to the period from 0 to 3 s following the first lick after reward delivery (i.e. the onset of reward consumption). Clusters 1 and 3 flip the sign of their reward responses from positive early in learning to negative late in learning. Early session was defined as the first day of learning and late session as the day when anticipatory licking in response to CS+ was high and stable (Materials and Methods) (Namboodiri et al., 2019). Error shadings correspond to confidence intervals.
- **D.** Schematic showing that reward responses late in learning is the sum of responses to reward receipt and reward prediction. The reward responses early in learning only reflect reward receipt as the animals have not yet learned to predict the reward.
- **E.** Schematic of a session in which reward probability is reduced to 50% after learning the CS+→reward association at 100% probability. A subtraction of responses between the rewarded and unrewarded trials isolates the reward receipt response late in learning.
- **F.** PSTH for cluster 1 showing that response on rewarded trials is lower than the response on unrewarded trials. In other words, the pure reward receipt response (difference between rewarded and unrewarded trials as shown in cyan) is negative after learning, even though it was positive early in learning (**C**).

G. Reward receipt responses early and late in learning measured using coefficients of a General Linear Model (GLM) fit. We performed the GLM analyses to remove lick response confounds (Namboodiri et al., 2019) even though consumption licking (unlike anticipatory licking) is comparable early and late in learning. Reward receipt responses late in learning were calculated by subtracting responses on rewarded and unrewarded trials in the 50% reward probability session. This is similar to what is shown in **E**, but with the difference that the GLM coefficients are calculated with respect to the first lick after reward delivery or reward omission. Red hexagons show the clusters with the largest reduction in reward receipt responses after learning. Even though cluster 5 appears to show a large difference in reward response between early and late in learning in **C**, this is because the reward prediction response increases late in learning. See Table S1 for statistical results for all recorded neurons and separately for only the neurons that were longitudinally tracked between these sessions. See Table S1 also for all statistical results in the manuscript, including all details. See **Fig S1** for responses of all neurons and clusters.

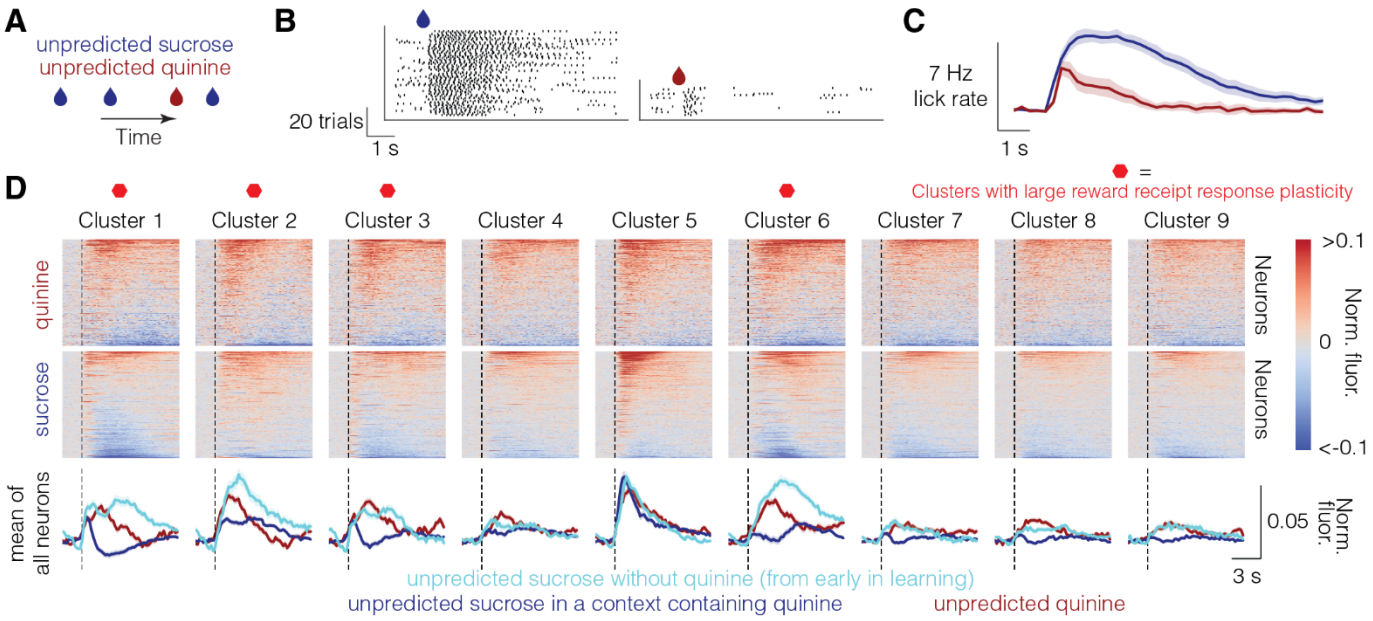


Fig 2: vmOFC unpredicted reward responses show similar plasticity in a context containing a highly salient aversive stimulus.

- **A.** Schematic of unpredicted sucrose and quinine delivery. Unpredicted sucrose (10%) and quinine (1.5-2.5 mM) are delivered pseudorandomly in a 3:1 ratio (Methods). In this experiment, licks are necessary to sample the liquid.
- **B.** Raster plot of licking (black ticks) from an example behavioral session from one animal. Animals lick at high rates after sucrose delivery, but immediately stop licking after sampling quinine deliveries.
- **C.** Average lick rate across all animals and sessions (n=26 sessions from n=11 OFC-CaMKII and OFC→VTA mice). The histograms are time-locked to liquid delivery.
- **D.** The average PSTH for sucrose and quinine responses for all OFC-CaMKII neurons, aligned to the first lick after liquid delivery (i.e. initiation of consumption). The bottom line graphs show the average across all neurons within a cluster. The cyan line shows the average response of each cluster to reward early in learning (same as in **Fig 1C** but aligned to first lick after liquid delivery). The red hexagons at the top show the clusters with the most reduction in sucrose responses due to the presence of quinine in the context. These are the same clusters as those showing suppression due to reward prediction (**Fig 1G**). In this case, since the reward response is evidently dissociated from licking (**Supplementary Note 2**), we did not perform a GLM analysis. Responses of OFC→VTA neurons are shown in **Fig S2**.

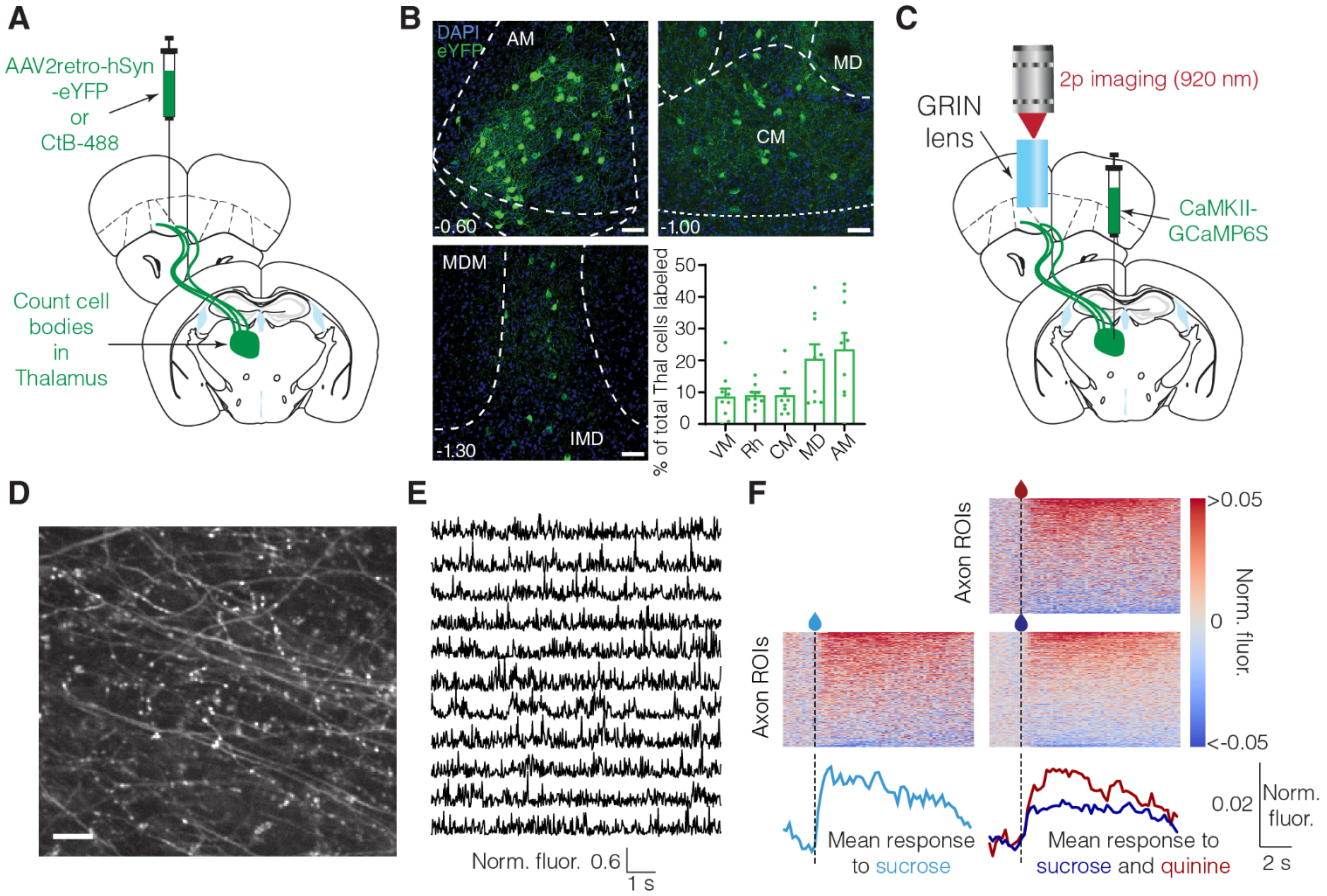


Fig 3: Medial thalamus (mThal) conveys reward responses to vmOFC and shows qualitatively similar plasticity as vmOFC neurons.

- **A.** Surgery schematic for retrograde anatomical tracing, showing injections of either retrogradely traveling virus (AAV2retro) or Cholera Toxin-B (CTB).
- **B.** mThal cell bodies projecting to OFC counted using CTB and AAV2retro labeling. Representative images show AAV2retro expression (see **Fig S3** for CTB expression). Top 5 thalamic regions are shown. See **Fig S3** for counts in all thalamic nuclei, split by AAV2retro and CTB injections. AM: Anteromedial, MD: Mediodorsal, CM: Centromedial, Rh: Rhomboid, VM: Ventromedial. Scale bar = 50 µm.
- **C.** Surgery schematic for mThal axon imaging in vmOFC.
- **D.** Example zoomed-in mThal axon standard deviation projection image showing individual axons in vmOFC. The scale bar corresponds to 10 µm.
- **E.** Example mThal GCaMP traces from individual axonal regions of interest (ROIs) (Methods).
- **F.** Heat maps show trial-averaged responses from individual axon ROIs (*do not necessarily correspond to distinct axons,* see **Fig S3** and Methods for details and interpretation) to unpredicted sucrose alone (left) or unpredicted sucrose and quinine (right, similar experiment as **Fig 2A**). The bottom traces show the average responses across all segmented mThal axon ROIs aligned to the first lick after liquid delivery (dashed line). The same animals (n=3) were used under all conditions, so as to be directly comparable; see **Fig S3** for data from two more animals in the sucrose only condition. See **Fig S3** also for quantification of heterogeneity of response patterns in mThal axons. Mean sucrose response across the population is lower in the session with quinine compared to the session without quinine. The statistical test was applied using the mean fluorescence across all ROIs per animal as an independent measure. This plasticity is qualitatively similar to that seen in some vmOFC clusters (**Fig 2**).

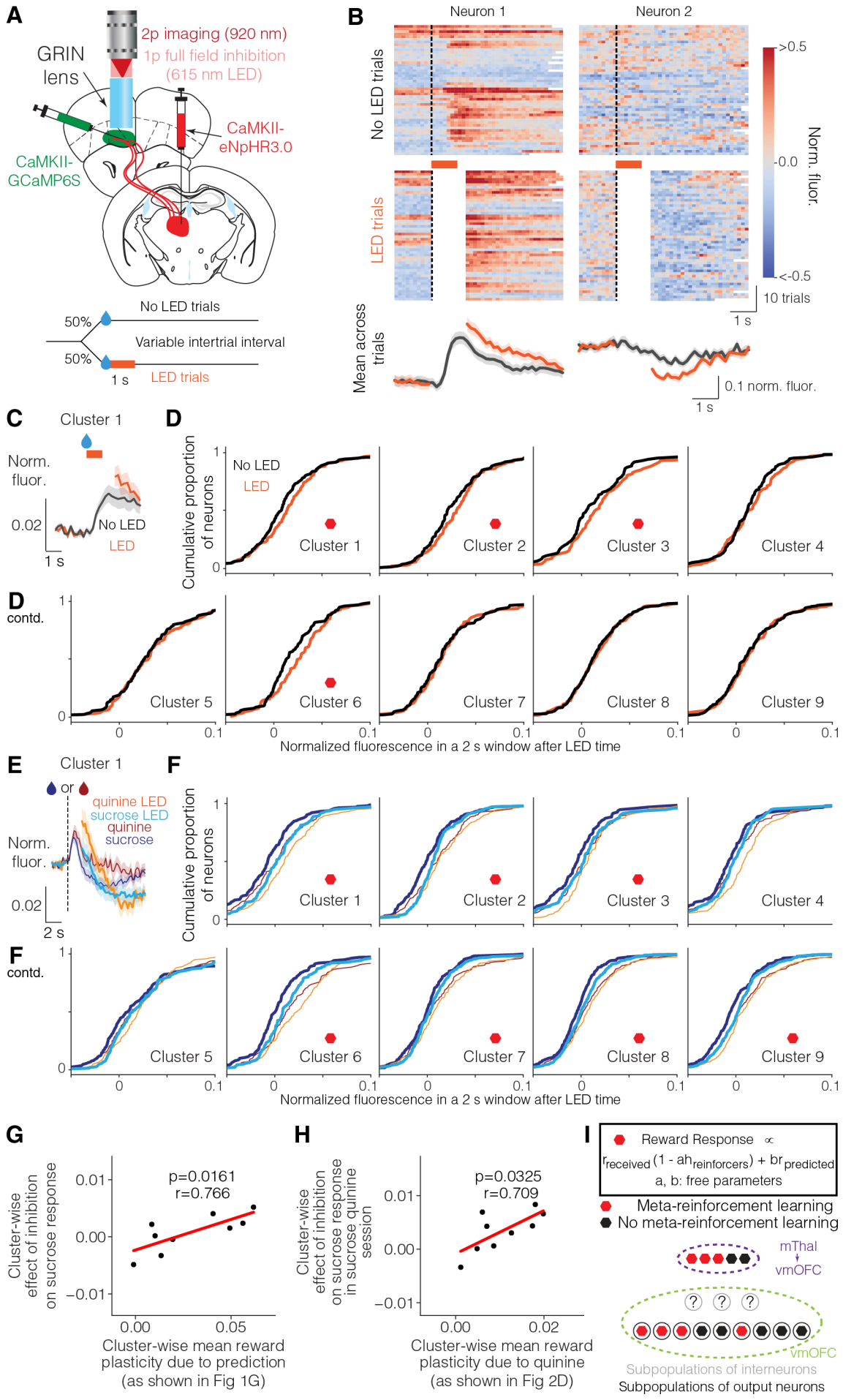

A
GRIN lens
2p imaging (920 nm)
1p full field inhibition (615 nm LED)
CaMKII-eNpHR3.0
CaMKII-GCaMP6S
No LED trials
50%
Variable intertrial interval
50%
1 s
LED trials
B
Neuron 1
Neuron 2
No LED trials
LED trials
>0.5
-0.0
<-0.5
Norm. fluor.
10 trials
1 s
Mean across trials
0.1 norm. fluor.
1 s
C
Cluster 1
Norm. fluor.
0.02
No LED
LED
1 s
D
Cumulative proportion of neurons
No LED
LED
Cluster 1
Cluster 2
Cluster 3
Cluster 4
D contd.
Cluster 5
Cluster 6
Cluster 7
Cluster 8
Cluster 9
Normalized fluorescence in a 2 s window after LED time
E
Cluster 1
or
quinine LED
sucrose LED
quinine
sucrose
Norm. fluor.
0.02
2 s
F
Cumulative proportion of neurons
Cluster 1
Cluster 2
Cluster 3
Cluster 4
F contd.
Cluster 5
Cluster 6
Cluster 7
Cluster 8
Cluster 9
Normalized fluorescence in a 2 s window after LED time
G
Cluster-wise effect of inhibition on sucrose response
p=0.0161
r=0.766
Cluster-wise mean reward plasticity due to prediction (as shown in Fig 1G)
H
Cluster-wise effect of inhibition on sucrose response in sucrose quinine session
p=0.0325
r=0.709
Cluster-wise mean reward plasticity due to quinine (as shown in Fig 2D)
I
Reward Response ∝
r_received (1 - ah_reinforcers) + br_predicted
a, b: free parameters
Meta-reinforcement learning
No meta-reinforcement learning
mThal
vmOFC
vmOFC
Subpopulations of interneurons
Subpopulations of output neurons

Fig 4: Inhibition of mThal reward responses disrupts vmOFC reward responses in a cluster-specific manner.

- **A.** Schematic of mThal inhibition while imaging vmOFC CaMKIIα expressing neurons. Top shows surgery schematic and bottom shows experiment schematic.
- **B.** Example neurons showing effect of mThal inhibition on unpredicted sucrose responses. Left neuron shows an increase in activity due to mThal inhibition, whereas the right neuron shows a decrease in activity. Frames around LED illumination were masked out (shown as white) to prevent light artifacts in imaging.
- **C.** PSTH of cluster 1 showing effect of mThal inhibition on unpredicted sucrose responses (no quinine or cues in the same session). The reward responses without LED are positive in all clusters, thereby providing an independent replication of positive reward responses early in learning (**Fig 2D**).
- **D.** Empirical cumulative distribution functions showing average fluorescence with and without LED for individual neurons within each cluster. This shows the full distribution of the population responses, with a rightward shift signifying an increase in activity. The red hexagons correspond to the clusters showing significant change in fluorescence due to inhibition. These same clusters show large changes in sucrose responses due to reward prediction (**Fig 1**) or presence of quinine (**Fig 2**). See **Fig S4** for controls.
- **E.** PSTH of cluster 1 showing effect of mThal inhibition on sucrose and quinine responses.
- **F.** Same as **D** for sucrose and quinine responses. Here, the red arrows corresponds to clusters showing significant mean effect on their sucrose responses. In addition to the four clusters above (**D**), clusters 7, 8 and 9 also show significant effect of mThal inhibition.
- **G.** Cluster-wise relationship between mean reward response plasticity due to reward prediction (drop in reward response due to prediction, **Fig 1G**) and effect of mThal inhibition on sucrose response (i.e. change in sucrose response due to LED as shown in **Fig 4D** minus the change in spontaneous response due to LED as shown in **Fig S4**). There is a strong positive correlation (~60% explained variance) implying that mThal input contributes the majority of the variability across clusters in reward response plasticity due to reward prediction.
- **H.** Cluster-wise relationship between mean reward response plasticity due to the presence of quinine (drop in reward response due to quinine, **Fig 2**) and effect of mThal inhibition on sucrose response (i.e. change in sucrose response due to LED as shown in **Fig 4F** minus the change in spontaneous response due to LED as shown in **Fig S4**). There is a strong positive correlation (~50% explained variance).
- **I.** Top: The simplest mathematical model for the reward response plasticity in vmOFC (extension of **Fig S1**). $r_{received}$ represents received reward, $r_{predicted}$ represents predicted reward and $h_{reinforcers}$ represents history of prior reinforcers such as cues predicting reward or presence of highly salient aversive stimuli such as quinine. $h_{reinforcers}$ is the term that negatively gates reward receipt response.

  Bottom: Schematic of circuit model showing different reward information streams from mThal to vmOFC, with some (red hexagons) showing reward response plasticity consistent with meta-reinforcement learning (**Fig S3**). The vmOFC local circuit contains distinct interneuron subpopulations (e.g. those expressing parvalbumin, somatostatin and vasoactive intestinal peptide) and distinct activity-labeled subpopulations of output neurons (i.e. the clusters in OFC-CaMKII neurons). Four output neuron clusters show plasticity consistent with meta-reinforcement learning. The connectivity between the different subpopulations shown (mThal,

vmOFC interneurons, vmOFC output neurons) remains to be worked out (see text).

STAR Methods

Subjects and Surgery:

All experimental procedures were approved by the Institutional Animal Care and Use Committee of the University of North Carolina and University of Washington, and accorded with the Guide for the Care and Use of Laboratory Animals (National Institutes of Health). Adult male and female wild type C57BL/6J mice (Jackson Laboratories, 6-8 weeks, 20-30 g) were group housed with littermates and acclimatized to the animal housing facility until surgery. Survival surgeries were stereotaxically performed while maintaining sterility, as described previously (Namboodiri et al., 2019; Resendez et al., 2016). Induction of anesthesia was carried out by using 5% isoflurane mixed with pure oxygen (1 L/min) for roughly thirty seconds to a minute, after which anesthesia was maintained using 0.6-1.5% isoflurane. The surgeon monitored respiratory rate intermittently to ensure appropriate depth of anesthesia. The animals were placed on a heating pad for thermal regulation. Data from animals used in **Fig 1**, **Fig 2**, **Fig S1** and 7 were collected at UNC. For these animals, pre-operative buprenorphine (0.1 mg/kg in saline, Buprenex) treatment was given for analgesia. Eyes were kept moist using an eye ointment (Akorn). 2% lidocaine gel (topical) or 1mg/kg lidocaine solution was applied or injected onto the scalp prior to incision. Details of viral injection, lens and optic fiber implantation are provided below. A custom made stainless steel ring (5 mm ID, 11 mm OD, 2-3 mm height) was implanted on the skull for headfixation and stabilized with skullscrews as well as dental cement. Animals either received acetaminophen (Tylenol, 1 mg/mL in water) in their drinking water for 3 days, or 5 mg/kg carpofen 30 min prior to termination of surgery for post-operative analgesia. Animals were given at least 21 days (and often, many more) with ad libitum access to food and water to recover from surgery. Following recovery, animals used for behavioral studies were water deprived to reach 85-90% of their pre-deprivation weight and maintained in a state of water deprivation for the duration of the behavioral experiments. Animals were weighed and handled daily to monitor their health. The amount of water given daily was between 0.6-1.2 mL and was varied for each animal based on the daily weight. A total of 29 mice were included in this study: 5 OFC-CaMKII imaging (0 female, **Fig 1, 2**), 7 OFC-VTA imaging (0 female, **Fig S1, 2**), 5 mThal→vmOFC axon imaging (3 female, **Fig 3, Fig S3**), 7 mThal→vmOFC inhibition or control inhibition during vmOFC imaging (3 female, Fig 4, **Fig S4**), 10 anatomical tracing (8 female, **Fig 3B**).

Head-fixed behavior:

Trace conditioning was done exactly as before (Namboodiri et al., 2019). A brief outline of these methods is summarized here. Water deprived mice were first trained to lick for random unpredictable sucrose (10-12.5%, ~2.5 μL) deliveries in a conditioning chamber. Mice received one of two possible auditory tones (3 kHz pulsing tone or 12 kHz constant tone, 75-80 dB) that lasted for 2 seconds. A second after the cues turned off, the mice received a sucrose reward following one of the tones (designated CS+), whereas the other tone resulted in no reward (designated CS-). The identity of the tones was counterbalanced across mice in all experiments. The cues were presented in a pseudorandom order and in equal proportion until a total of 100 cue presentations (trials) were completed. The intertrial interval between two consecutive presentations of the cues was drawn from a truncated exponential distribution with mean of 30 s and a maximum of 90 s, with an additional 6 s constant delay. Early in learning (**Fig 1**) was defined as the first session of conditioning. Late in learning (**Fig 1**) was defined as the day that the area under a receiver operating characteristic curve (auROC) of lick rates to CS+ versus CS- remained high and stable (auROC larger than 0.7 on at least 2 consecutive sessions or larger than 0.85).

For the sucrose and quinine experiment, drops of sucrose (10-12.5%, ~2.5 μL) or quinine hydrochloride dihydrate (1.5-2.5 μM, ~2.5 μL) were randomly delivered at a 3:1 ratio (60 sucrose drops and 20 quinine drops). The interdrop interval was a minimum of 13-18 s and a maximum of 23-28 s. These were chosen to maintain a sufficient interval between consecutive drops so as to prevent any bleed-through of GCaMP fluorescence. Even though the hazard rate is not flat for these intervals, the animals did not show any behavioral evidence of temporal expectation of the delivery times.

2-photon microscopy:

The methods were similar to those published previously (Namboodiri et al., 2019). We used a calcium indicator (GCaMP6s) to image calcium changes using 2-photon microscopy. The injection coordinates and volumes for virus as well as the coordinates for implanting a gradient refractive index (GRIN) lens were as published previously (Namboodiri et al., 2019). For mThal axon imaging (**Fig 3**) or inhibition (**Fig 4**), we injected 400-500 nL of AAVDJ-CaMKIIα-GCaMP6s (at an effective titer of ~1-2×10^{12} infectious units per mL, UNC Vector Core) or AAV5-CaMKII-eNpHR3.0-mCherry (~4×10^{12} infectious units per mL, UNC Vector Core) unilaterally in mThal (-1.3 AP, 0.5 ML, -3.5 DV from bregma). We used a resonant scanner (30 Hz frame rate acquisition, Olympus Fluoview FVMPE-RS) and performed an online averaging of 6 times to get an effective frame rate of 5 Hz, to minimize the size of recorded files as we had negligible motion artifacts. A GaAsP-PMT with adjustable voltage, gain and offset was used, along with a green filter cube. We used a long working distance 20x air objective, specifically optimized for infrared wavelengths (Olympus, LCPLN20XIR, 0.45 NA, 8.3 mm WD). We imaged either at 955 or 920 nm using a Ti-sapphire laser (SpectraPhysics, ~100 fs pulse width) with automated alignment. The animals were placed on a 3-axis rotating stage to precisely align the surface of the GRIN lens to be perpendicular to the light path, such that the entire circumference of the lens is crisply in focus (within 1-2 μm). The imaging acquisition was triggered by a custom Arduino code right before the start of a behavioral session, and a TTL output of every frame was sent as an input to the Arduino. The imaging acquisition was triggered off at the end of the behavioral session (~ one hour). The data in **Fig 1** (OFC-CaMKII) and **Fig S1** (OFC→VTA) were re-analyzed based on data from a previous publication (Namboodiri et al., 2019). In every mouse, one z-plane was imaged throughout acquisition so that the same cells could be tracked through learning (Namboodiri et al., 2019). After mice were trained, other z-planes were also imaged (one per session) to get a measure of the total functional heterogeneity in the network. A total of 2-6 z-planes per mouse were imaged in the OFC-CaMKII group, whereas 1-3 z-planes were imaged in the OFC-VTA group. Thus, responses early in learning were from the plane tracked throughout learning, whereas the responses late in learning were from all imaged planes. The sucrose quinine session was run for most imaging planes at the end of the conditioning experiments (Namboodiri et al., 2019). The data in **Fig 2** were collected from the same animals as in **Fig 1**, but have never been published previously.

Imaging data analysis:

Preprocessing (motion correction, manual ROI detection, signal extraction, neuropil correction) using SIMA (Kaifosh et al., 2014) for vmOFC cell body imaging was as described previously (Namboodiri et al., 2019). For axon imaging, we used the same approach as for cell bodies for motion correction. For detection of axonal ROIs, we employed a manual hand-drawing method but with different criteria: 1) we drew ROIs only around parts of an axon that showed no resolvable overlap with other fluorescent regions, thereby making the ROIs small, 2) we drew ROIs along a single axon (at least those that can reliably be tracked along the imaging focal plane) only once (more on this below), 3) we drew ROIs

around regions of an axon that are definitely within the imaging plane (these are often bright and sharply in focus), 4) we drew ROIs that were at least 10 pixels or so, in order to minimize noise. An example plane with manually annotated ROIs is shown in **Fig S3**. Despite these precautions, it is impossible to know whether different ROIs are from the same underlying axon, as axons move in and out of the imaging plane and are highly branched. Thus, we do not make any claims about the individual ROIs shown in **Fig 3F** representing individual axons (though the example traces in **Fig 3E**, representing the axonal ROIs with the largest skew in activity, show distinct patterns). A common approach to identify ROIs that are putatively from the same axon is to remove segments that show high correlations. However, this threshold depends crucially on the signal to noise ratio of the recording. If the signal to noise ratio is low, segments of the same axon will show low correlations due to the noise dominating the fluorescence. Thus, the decision to set the threshold often becomes subjective, especially considering variability in signal to noise ratio between animals. We avoid this problem by not claiming that different ROIs are necessarily from different axons. Instead, we quantify both a population mean of all ROIs and separately perform a clustering analysis of ROIs based on their response profiles to identify heterogeneous response profiles (more on this below). We did not perform any neuropil correction for axon recordings as the neuropil is primarily due to signals of interest (i.e. from axons). Due to this reason, our earlier approach of ensuring no overlap with other axons is important to get isolation of signals.

The clustering analysis for vmOFC neurons was performed as previously published (Namboodiri et al., 2019). Importantly, all clustering was performed based on the peristimulus time histograms (PSTHs) around CS+ and CS- in the session after learning with 100% reward probability. So, any neuron that has an assigned cluster identity was recorded under the 100% contingency late in learning. Once a cluster ID was assigned to a neuron, the same cluster ID was used for that same neuron in all other sessions. This was possible because 2-photon imaging allowed longitudinal tracking of the exact same set of neurons across many days and tasks (Namboodiri et al., 2019).

The clustering of axon ROIs (**Fig S3**) used the same approach but used PSTHs around sucrose alone (**Fig S3G**), or around both sucrose and quinine (**Fig S3H**). As we did not longitudinally track axons across these sessions, we performed clustering separately for these two sessions. Nevertheless, qualitative correspondence between the sucrose responses of some clusters can be seen by comparing **Fig S3G** and **H**. The benefit of performing cluster analysis for axons is that different identified clusters are almost certain to be from different underlying axons, as the clustering is done on average responses across trials, thereby reducing noise. Thus, we do not assess axon-by-axon heterogeneity as we cannot reliably identify individual axons. We can nevertheless assess heterogeneity of information encoding in the axonal population by interpreting cluster-wise differences in response profiles. This is philosophically similar to our approach with clustering of vmOFC neurons. We only interpret average results across all neurons within a cluster, thereby treating a cluster as a unit of information representation. The one big difference between the identification of cell body and axon clusters is that unlike in the cell body case, we cannot assess prevalence of each cluster among the axonal population. This is because the ROIs making up any cluster might potentially be overlapping and represent the same axon.

In **Fig 1**, we quantified reward responses as the coefficients of a General Linear Model (GLM) fit to reward delivery (Namboodiri et al., 2019). Importantly, this GLM approach was applied to the deconvolved calcium fluorescence so as to putatively remove fluorescence changes purely due to the dynamics of GCaMP6s. We employed a GLM approach primarily to separate lick related responses and reward prediction or receipt responses, as both licking and rewards generally produced positive responses (Namboodiri et al., 2019). We did not employ a similar GLM approach for analyzing the

sucrose and quinine responses in **Fig 2** as these responses were evidently dissociated from licking responses. This is because licking was high for sucrose alone, high for sucrose in the sucrose-quinine session and low for quinine in the sucrose-quinine session; yet, the responses were generally high, low and high respectively, for these conditions.

Optogenetics during imaging:

These animals received an injection of AAVDJ-CaMKIIα-GCaMP6s in vmOFC and AAV5-CaMKIIα-eNpHR3.0-mCherry (experimental) or AAV5-CaMKIIα-mCherry (control) in mThal. All animals showed significant expression of the opsin in mThal. Light was delivered for optogenetic inhibition to the full field of imaging by an LED kit with a peak wavelength of 615 nm (FV30SP-LED615, Olympus) (Otis et al., 2019). The frames containing light artifacts due to LED illumination were masked out for analysis (shown as white bars in **Fig 4B, C, E**). The same preprocessing pipeline as before (including motion correction, signal extraction and neuropil correction) was employed on this masked data. These animals were first trained to lick in response to random sucrose deliveries. We then performed optogenetic inhibition during sucrose consumption for multiple imaging planes (**Fig 4 B-D**). A random half of the trials received inhibition. The effect of the LED was calculated by comparing sucrose fluorescence on the trials with and without LED. The animals were subsequently trained on the trace conditioning paradigm (no imaging). Once anticipatory licking was high and stable, we imaged the same neurons that were imaged earlier so as to obtain PSTHs around CS+ and CS-. These PSTHs were used to classify neurons into the clusters identified using the much larger population of neurons in **Fig 1**. The classification was done using a linear support vector classifier (Scikitlearn), as was used previously for classifying OFC→VTA neurons (Namboodiri et al., 2019). We then performed inhibition of mThal axons while imaging from vmOFC neurons during the sucrose and quinine session. Here, in order to obtain a sufficient number of trials with and without inhibition, we first performed recordings without inhibition (80 trials), followed by with inhibition (80 trials). The effect of LED was calculated by comparing sucrose or quinine fluorescence with and without LED.

Retrograde tracing, histology and microscopy:

400 nL of rAAV2retro-hSyn-eYFP (~$2x10^{12}$ infectious units/mL) or CTB-488 were stereotaxically injected at roughly 2.5 AP, 0.5 ML and 2.3-2.5 DV from bregma using the surgical methods described above. 3-5 weeks after surgery, animals were euthanized with an overdose of pentobarbital (~390 mg/kg, Somnasol, Covetrus EU-HS-045-100-0), and transcardially perfused with 4% paraformaldehyde (PFA, Sigma-Aldrich, #158127). Perfused brains were incubated in 4% PFA overnight and moved to a 30% sucrose solution (Sigma-Aldrich, #S0389) for ~2 days prior to cryosectioning. 40 µm thick sections were used in tracing experiments. For retroAAV2 thalamic labeling (**Fig 3B, Fig S3**), eYFP signal was enhanced and stabilized using a chicken anti-GFP antibody (Aves Lab, #GFP-1020, 1:500 dilution), paired with a donkey anti-chicken secondary (Jackson Immunoresearch, #703-545-155, 1:1000 dilution). GFP and eYFP have highly similar protein sequences, which allows the use of a GFP antibody for immunostaining. Brain sections were imaged using a 20x air objective on a confocal microscope (Olympus Fluoview FV3000). Resulting image tiles were stitched, and Z-stacks were taken at ~1 µm intervals and averaged across slices yielding a maximum intensity projection image. Brain atlas outlines (https://mouse.brain-map.org/static/atlas) were overlaid onto each image to allow assignment of thalamic subregions, in which labeled cells were counted using ImageJ (https://imagej.net/Fiji). Percentage of total thalamic cells labeled (**Fig 3B**) was quantified as the number of cells per region divided by the sum total of all counted eYFP+ or CTB+ cells. The intermediodorsal nucleus was counted as part of the mediodorsal region.

Supplementary Information

Supplementary Notes

1. Our hypothesis that the presence of quinine makes sucrose comparatively less salient for learning, hence reducing vmOFC sucrose responses, was based on three converging reasons. One, when directly compared, punishments/salient aversive stimuli typically have higher learning rates, as discussed in the main text. Two, the mice in our task had considerable experience receiving sucrose under the two-photon microscope. However, quinine was a novel stimulus in the sucrose-quinine session. Novelty typically increases salience. Three, in addition to learning the CS+-reward association, mice also learn a context-reward association (as mentioned in point two above). The presence of quinine reduces the context-reward association. Hence, learning about the frequency of quinine is more important than learning about the frequency of sucrose: if the quinine frequency is high, mice should stop licking altogether to avoid exposure. We verified that this is indeed the case (data not shown). All three of the above reasons imply that sucrose is relatively less salient for learning in the presence of quinine. Our hypothesis, findings and interpretation are independent of the precise combination of the above factors contributing to a reduction in salience for sucrose in the presence of quinine.

2. The reduction in sucrose responses in the presence of quinine cannot be explained by a difference in licking behavior to sucrose and quinine. This is because the consumption lick rate is high for unpredicted sucrose without quinine, high for unpredicted sucrose in a context containing quinine, and low for unpredicted quinine. However, the responses are high for unpredicted sucrose without quinine, low for unpredicted sucrose in a context containing quinine, and high for unpredicted quinine.

3. One potential confound for the plasticity observed in vmOFC sucrose responses due to the presence of quinine is that there might be some interaction between the lingering taste of a previous quinine drop and a sucrose drop. However, this cannot explain the results we observed. This is because quinine produced a positive response much like what sucrose produced when delivered alone (**Fig 2**). If the taste of sucrose and quinine interact, it should produce a sucrose response similar to that of the quinine response. Instead, we observed that the sucrose response diverged from the quinine response, a finding that cannot be explained by simple history effects of taste.

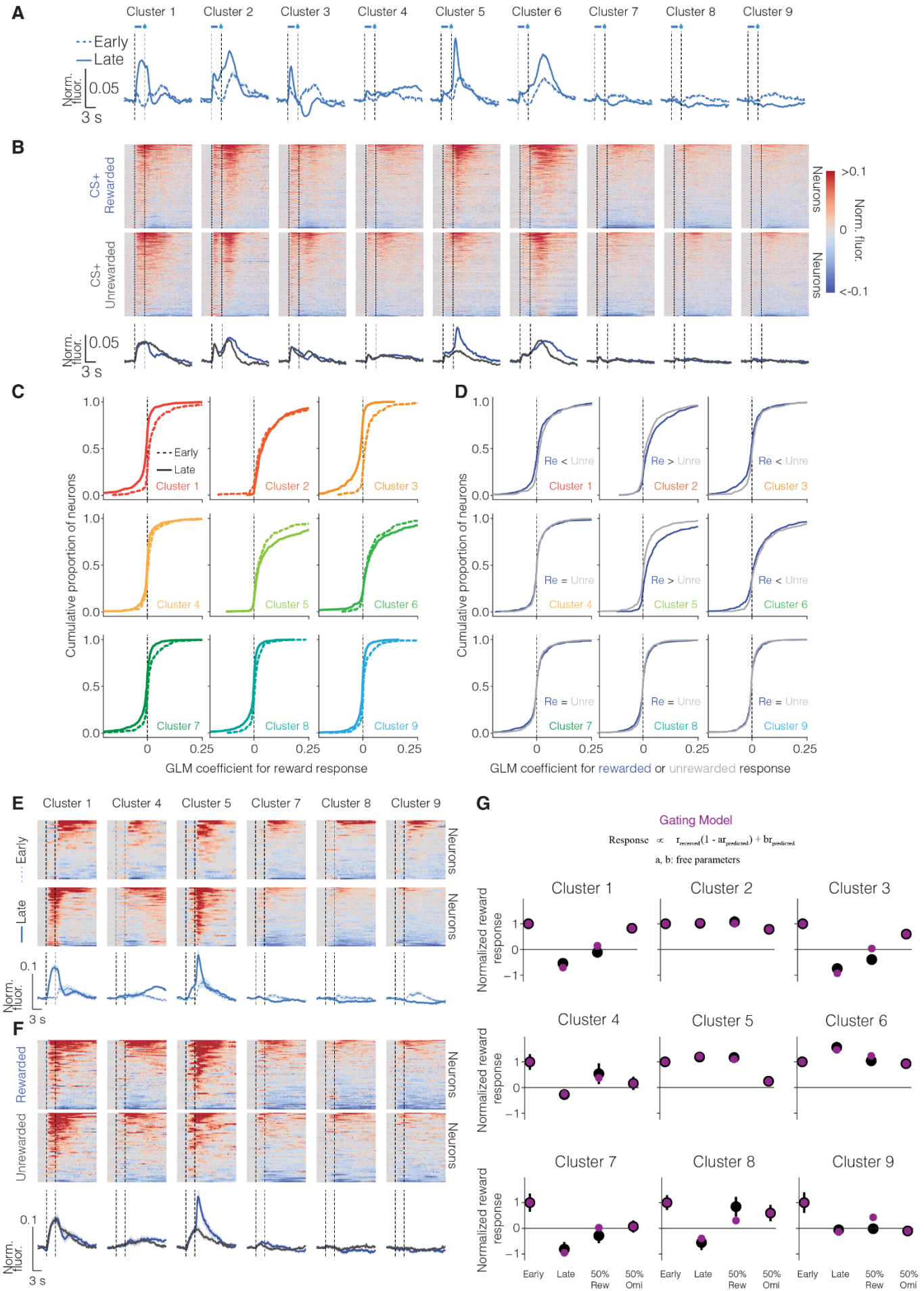

A
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Cluster 8
Cluster 9
Early
Late
Norm. fluor.
0.05
3 s
B
CS+ Rewarded
CS+ Unrewarded
Neurons
>0.1
0
<-0.1
Norm. fluor.
C
Cumulative proportion of neurons
GLM coefficient for reward response
D
Re < Unre
Re > Unre
Re = Unre
GLM coefficient for rewarded or unrewarded response
E
F
Rewarded
Unrewarded
G
Gating Model
Response ∝ r_received(1 - ar_predicted) + br_predicted
a, b: free parameters
Normalized reward response
Early
Late
50% Rew
50% Omi

Fig S1 (related to Fig 1): Reward responses of all OFC-CaMKII and OFC→VTA neurons

- **A.** PSTHs of all OFC-CaMKII clusters early and late in learning. These data were published previously, but focused on analysis of cue responses (Namboodiri et al., 2019). The reward responses early in learning are temporally smeared since the latency to consume the reward after delivery (i.e. first lick after reward delivery) is often variable on this first session. Therefore, reward receipt responses were always analyzed time-locked to the actual consumption starting on the first lick after reward delivery.
- **B.** PSTHs of all OFC-CaMKII neurons on rewarded and unrewarded trials in the 50% reward probability session. The same ordering of neurons across rows is maintained for rewarded and unrewarded trials. Mean of all neurons within a cluster are shown in the line traces at the bottom. See number of neurons per cluster in the row for panel **D** in Table S1. Late reward responses (3-6 s after consumption) (Namboodiri et al., 2019) are not analyzed in this study and are positive for clusters 2 and 6.
- **C.** Empirical cumulative distribution function of reward responses (measured using GLM coefficient) early and late in learning for all neurons in all clusters. The limits of the x axis were set to be the same across all graphs for comparison. Part of the positive tail for some clusters is cropped due to this standardization. All neurons recorded per condition are shown. See **Fig S3** for data only from neurons longitudinally tracked across all conditions.
- **D.** Empirical cumulative distribution function of responses on rewarded and unrewarded trials. The comparison of the mean response on rewarded and unrewarded trials is shown for each cluster as greater than, less than or equal.
- **E.** The PSTHs and averages per cluster of all OFC→VTA neurons are shown for sessions early and late in learning. The color bar is the same as in **B**.
- **F.** Same as above for 50% probability session showing average responses for rewarded and unrewarded trials. These results show qualitative correspondence with the corresponding OFC-CaMKII clusters. The color bar is the same as in **B**.
- **G.** Results of gating model fits for all longitudinally tracked neurons (n=1572) across all four conditions for each cluster. Reward responses were measured by GLM coefficients. The reward receipts and predictions were set as (1, 1, 1, 0) and (0, 1, 0.5, 0.5) for the early, late, 50% reward, 50% omission conditions. Since only four experimental conditions were performed, we did not attempt a formal model comparison approach. The primary claim is that the non-linear gating model is the simplest model consistent with these data.

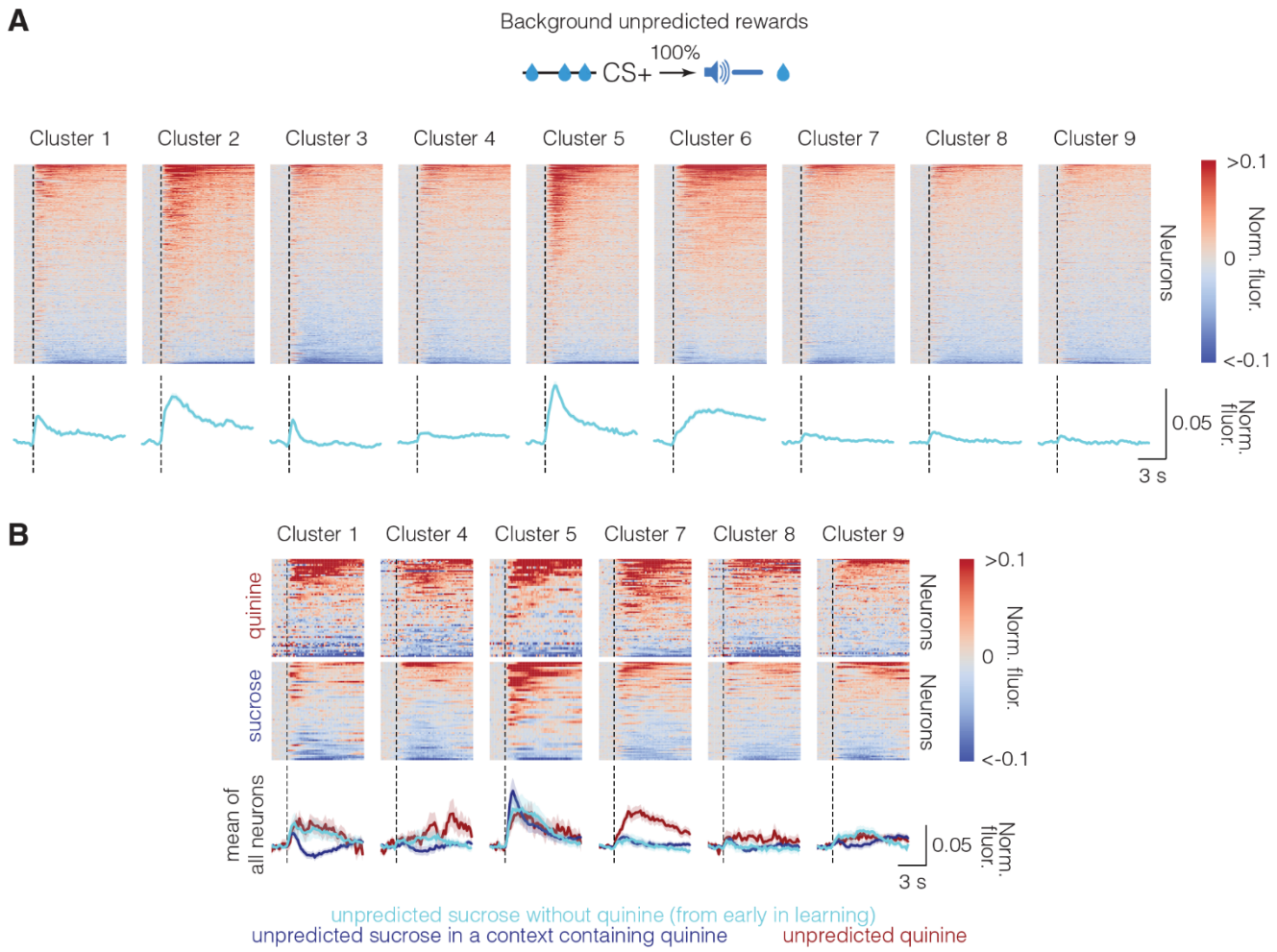
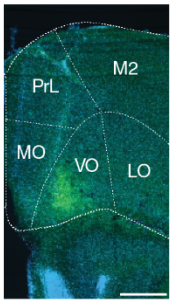
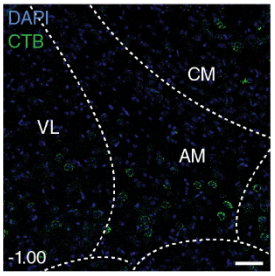


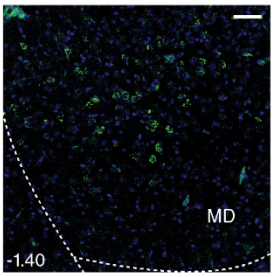
Fig S2 (related to Fig 2): A. Unpredicted reward responses after learning. We previously conducted a contingency degradation experiment after the learning of a 100% CS+→reward contingency (labeled "Background") (Namboodiri et al., 2019). The schematic of this experiment is shown at the top. The bottom graphs plot the responses of all neurons to the unpredicted rewards in the intertrial interval. An important caveat for this experiment is that the unpredicted rewards were often delivered at a high frequency (exponentially distributed with minimum delay between consecutive deliveries of zero; maximum average frequency of reward delivery was 1/6 Hz), potentially causing interactions between consecutive deliveries. **B.** PSTH from all OFC→VTA neurons during sucrose and quinine, with responses from early in learning overlaid. Similar plot as in **Fig 2D** for OFC→VTA neurons.

A

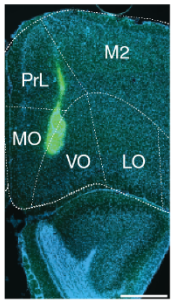


B





C



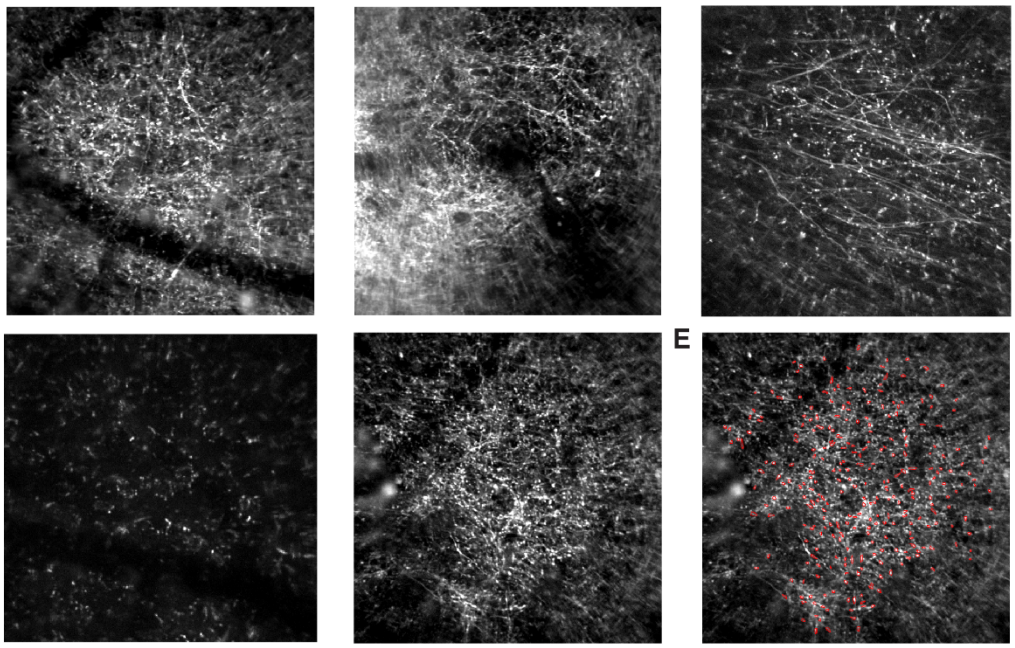
% of total eYFP labeling across all animals

	a1	a2	a3	a4	a5
aPVT	6.0	3.3	7.6	3.4	1.9
pPVT	7.3	10.9	2.0	5.0	0.7
PT	1.9	0.0	1.3	2.3	0.0
Re	9.2	2.5	10.2	4.4	11.6
CM	4.7	16.3	9.5	7.4	7.5
AM	38.3	25.5	26.3	44.0	41.8
MD	10.2	20.9	6.6	7.0	6.8
VAL	3.5	0.0	2.0	0.0	2.4
IAM	3.1	7.9	3.9	9.7	2.7
VM	6.8	3.3	13.5	9.4	10.1
Sub	1.6	1.3	7.2	1.3	6.8
Rh	7.3	7.9	9.9	6.0	7.7

% of total CTB labeling across all animals

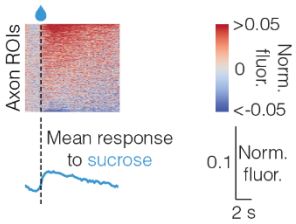
	a1	a2	a3	a4
aPVT	16.1	1.3	3.3	0.0
pPVT	4.3	1.3	1.6	3.8
PT	0.5	0.0	0.0	0.0
Re	3.8	8.1	1.6	7.7
CM	5.9	3.1	3.3	23.1
AM	8.6	8.8	13.9	0.0
MD	33.3	28.8	19.7	46.2
VAL	0.0	9.4	46.7	0.0
IAM	3.8	1.9	4.9	0.0
VM	7.0	25.6	0.8	0.0
Sub	4.3	1.9	0.0	3.8
Rh	12.4	10.0	4.1	15.4

D

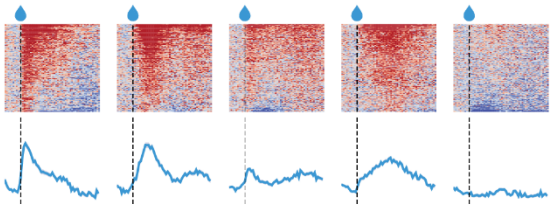


E

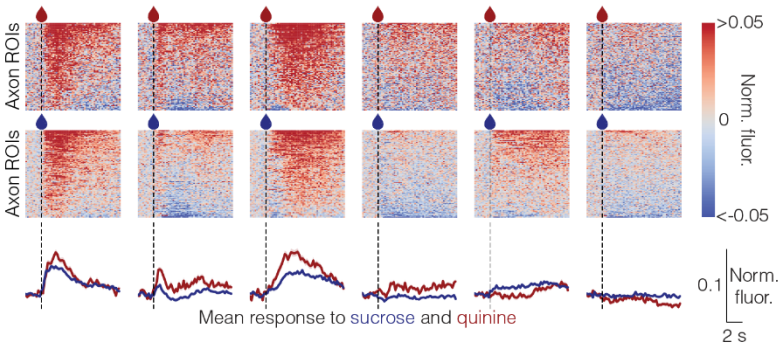
F



G



H



I

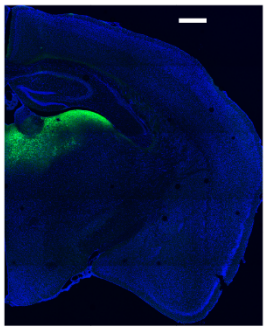


Fig S3 (related to Fig 3): rAAV2retro and CTB injection sites, signal quantification, and heterogeneity in mThal→vmOFC axonal reward responses.

- **A.** (Top) Example injection site for rAAV2retro injection into the OFC for anatomical tracing of inputs to vmOFC. Scale bar = 500 μm. Blue = DAPI, green = eYFP. (Bottom) Table illustrating the breakdown of the percent of eYFP+ cells in each thalamic subregion of each animal (animals a1-a5). AM: Anteromedial, aPVT: Anterior paraventricular thalamus, Cg2: cingulate cortex area 2, CM: Centromedial, IAM: Interanteromedial, LO: lateral orbitofrontal cortex, M2: secondary motor cortex, MD: Mediodorsal, MO: medial orbitofrontal cortex, pPVT: Posterior Paraventricular thalamus, PrL: prelimbic cortex, PT: parataenial nucleus, Re: Nucleus of reuniens, Rh: Rhomboid nucleus, Sub: Submedial nucleus, VAL: Ventral anterior-lateral, VM: Ventromedial, VO: ventral orbitofrontal cortex.
- **B.** Representative images for CTB-488 labeling in the thalamus and quantification of labeling in animals receiving CTB-488 injection in the vmOFC. Scale bar 50 μm.
- **C.** Same as in **A** for CTB-488 injections. Since the difference between CTB and rAAV2retro labeling could be either due to viral tropism or due to slight differences in injection sites, we decided to pool these together to create an average map of mThal input regions to vmOFC in **Fig 3B**.
- **D.** Standard deviation projection images of activity from the entire fields of view (~270 μm) of five different animals with mThal→vmOFC axons. Brightness and contrast were adjusted to maximize visibility.
- **E.** Manually drawn ROIs for individual axon segments shown in red. We took care to not double count different ROIs from the same axon (i.e. axons that extend within the imaging plane). However, since axons branch in and out of the imaging focal plane, it is possible that different axonal ROIs actually derive from the same axon. We did not attempt to classify ROIs as resulting from a single axon using a threshold of activity correlation since such measures depend crucially on the signal to noise ratio of the recording; for low signal to noise ratio recordings, two segments from the exact same axon can have low correlation. Thus, we instead only make claims about the population activity, either as the average, or as the average of an activity-defined cluster (**G**, **H**).
- **F.** Same as in **Fig 3F**, but including all five animals. The sucrose quinine experiment could not be run in the first two animals shown in the top row.
- **G.** Results of clustering axonal activity patterns (same approach as used for vmOFC cell bodies) showing five clusters of axon ROIs with distinct response profiles to sucrose. This analysis shows that distinct reward response profiles exist within mThal→vmOFC axons. The benefit of analyzing using a clustering approach is that we know that axon ROIs that are potentially from the same axons have to necessarily be within the same cluster since they will show the same response pattern (i.e. signal is the same irrespective of signal to noise ratio). Thus, we only interpret the results at the level of a cluster and state that these distinct response profiles exist within mThal→vmOFC axons. An important caveat is that we cannot make any claims regarding the relative prevalence of each response type or cluster due to the potential of double counting of axons within ROIs.
- **H.** Same as in **G** but for sucrose and quinine responses. The clustering was done separately for all axon ROIs recorded in this session as they were not longitudinally tracked from the sucrose only session. The results qualitatively match response patterns in vmOFC since 4 out of 6 clusters show larger responses for quinine compared to sucrose. Interestingly, two clusters (right most) appear to have larger responses for sucrose compared to quinine, potentially signaling value or valence of the liquid.
- **I.** Representative histological image showing expression of GCaMP6s in mThal cell bodies. Though there is relatively widespread expression in thalamus, only axons projecting to vmOFC were imaged (shown in **D**). Scale bar is 500 μm.

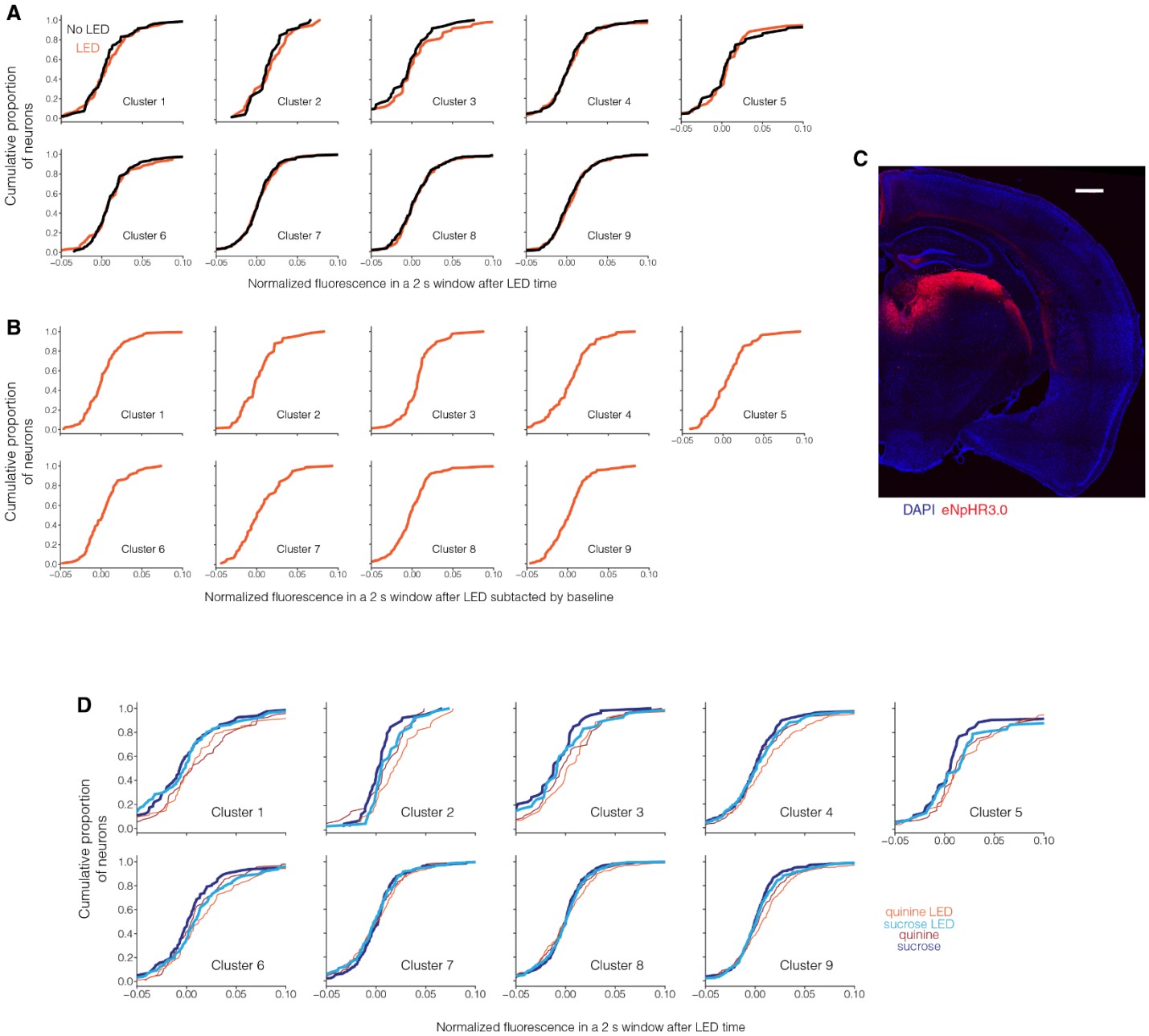


Fig S4 (related to Fig 4): Controls for mThal→vmOFC axon inhibition.

- **A.** Empirical cumulative distribution function of vmOFC neuronal responses to unpredicted sucrose (no quinine) with and without LED in control animals expressing a fluorophore (mCherry) without eNpHR3.0. No cluster had a mean LED response significantly different from zero.
- **B.** Empirical cumulative distribution function of vmOFC neuronal responses with LED in experimental animals expressing eNpHR3.0, but in the absence of any rewards (i.e. spontaneous inhibition of inputs). The responses were compared to the baseline right before inhibition. No cluster had a mean LED response significantly different from zero.
- **C.** Representative histological image showing expression of eNpHR3.0-mCherry in mThal cell bodies. Though there is relatively widespread expression in thalamus, only axons projecting to vmOFC were inhibited. Scale bar is 500 μm.
- **D.** Empirical cumulative distribution function of vmOFC neuronal responses to sucrose and quinine with and without LED in control animals expressing a fluorophore (mCherry) without eNpHR3.0.

Clusters 5 and 6 had a mean LED response significantly different from zero for sucrose, and clusters 2, 4, 6 and 9 had a mean LED response significantly different from zero for quinine. This suggests that there is a non-selective change in quinine responses in some clusters (as also seen in clusters 1 and 4 in **Fig 4**). The sucrose response changed due to LED in 7 out of 9 clusters in experimental animals and 2 out of 9 in the controls, likely suggesting that most of it is a selective change due to inhibition of the inputs. A direct comparison between the two groups is not possible since the expression levels of GCaMP6s and eNpHR3.0 vary across animals.

Table S1 (Statistical Results)

Figure	Description	Test	Statistic	p value	Number of samples	Multiple comparisons correction
Fig 1G	Difference in GLM coefficients between Late and Early	Two sample t test (treated as independent here; see next row for analysis of only the longitudinally tracked neurons between these sessions)	t for all clusters = (-5.74, -4.35, -6.35, -2.29, 0.07, -6.00, -2.47, -2.64, -2.38)	Two-tailed p values per cluster = (5.09 x10^{-08}, 6.81 x10^{-05}, 7.48 x10^{-09}, 2.54 x10^{-02}, 9.45 x10^{-01}, 1.55 x10^{-08}, 2.08 x10^{-02}, 1.51 x10^{-02}, 2.30 x10^{-02})	Number of neurons per cluster Late = (470, 458, 497, 572, 416, 423, 596, 693, 638) Early = (202, 96, 137, 150, 190, 179, 201, 245, 190)	BH correction across clusters
Fig 1G	Difference in GLM coefficients between Late and Early for longitudinally tracked neurons between these sessions	Paired t test	t for all clusters = (-6.44, -4.42, -5.23, -1.10, -0.32, -4.05, -2.89, -1.61, -2.00)	Two-tailed p values per cluster = (7.91 x10^{-09}, 8.05 x10^{-05}, 2.85 x10^{-06}, 3.08 x10^{-01}, 7.53 x10^{-01}, 1.75 x10^{-04}, 7.69 x10^{-03}, 1.40 x10^{-01}, 7.12 x10^{-02})	Number of neurons per cluster = (200, 95, 137, 147, 189, 179, 196, 243, 186)	BH correction across clusters
Fig 2C	Mean lick rate over 0-3 s following deliver of sucrose versus quinine	Paired t test	t(25) = 9.13	One-tailed p value = 9.7x10^{-10}	Number of sessions = 26, Number of animals = 11	NA
Fig 2D	Difference in mean fluorescence (0-3 s) between sucrose with quinine and sucrose without quinine	Two sample t test (treated as independent here; see next row for analysis of only the longitudinally tracked neurons	t for all clusters = (-6.61, -5.19, -5.26, -2.23, -0.31, -7.39, -4.43, -4.20, -3.74)	Two-tailed p values per cluster = (6.19 x10^{-10}, 1.89 x10^{-06}, 8.38 x10^{-07}, 2.98 x10^{-02}, 7.56 x10^{-01}, 7.41 x10^{-12}, 2.35 x10^{-05},	Number of neurons per cluster sucrose with quinine = (379, 331, 337, 421, 367, 352, 508, 527, 494)	BH correction across clusters

		between these sessions)		4.67 x10^{-05}, 2.89 x10^{-04})	Sucrose without quinine = (202, 96, 137, 150, 190, 179, 201, 245, 190)	
Fig 2D	Difference in mean fluorescence (0-3 s) between sucrose with quinine and sucrose without quinine for longitudinally tracked neurons between these sessions	Paired t test	t for all clusters = (-12.40, -6.02, -6.70, -2.60, -1.19, -8.23, -5.53, -5.21, -4.91)	Two-tailed p values per cluster = (1.77 x10^{-25}, 7.55 x10^{-08}, 1.57 x10^{-09}, 1.16 x10^{-02}, 2.35 x10^{-01}, 1.92 x10^{-13}, 1.77 x10^{-07}, 6.19 x10^{-07}, 2.52 x10^{-06})	Number of neurons per cluster = (197, 95, 135, 149, 189, 176, 200, 242, 188)	BH correction across clusters
Fig 3F	Difference in mean fluorescence (0-3 s) between sucrose alone versus sucrose in sucrose-quinine	Paired t test. In this case, we used each individual animal as an independent sample since we are taking the average fluorescence across all axon ROIs per animal	t(2) = -5.33 Norm. mean fluor. difference for the 3 animals between the sessions with and without quinine = (-0.0078, -0.0045, -0.0089)	Two-tailed p value = 0.033	Number of animals = 3	NA
Fig 4D	Mean difference in fluorescence response to sucrose with and without LED in a 2 s epoch	Paired t test	t for all clusters = (2.82, 2.56, 3.67, 1.68, 0.63, 3.30, 0.75,	Two-tailed p values per cluster = (1.65 x10^{-02}, 2.61 x10^{-02}, 3.81 x10^{-03}, 1.71 x10^{-01}, 5.33 x10^{-01}, 5.99 x10^{-03},	Number of neurons per cluster = (147, 139, 88, 133, 125, 100, 145, 259, 229)	BH correction across clusters

			0.83, 0.75)	5.10 x10^{-01}, 5.10 x10^{-01}, 5.10 x10^{-01})		
Fig 4F	Mean difference in fluorescence response to sucrose with and without LED in a 2 s epoch	Paired t test	t for all clusters = (5.11, 4.18, 3.58, 1.64, 1.44, 4.01, 2.50, 5.25, 4.35)	Two-tailed p values per cluster = (8.13 x10^{-06}, 2.08 x10^{-04}, 1.24 x10^{-03}, 1.57 x10^{-01}, 1.99 x10^{-01}, 2.93 x10^{-04}, 2.70 x10^{-02}, 4.95 x10^{-06}, 1.15 x10^{-04})	Number of neurons per cluster = (160, 178, 95, 151, 147, 141, 176, 317, 280)	BH correction across clusters and conditions (i.e. tests were done for both sucrose and quinine)
Fig 4F	Mean difference in fluorescence response to quinine with and without LED in a 2 s epoch	Paired t test	t for all clusters = (4.06, 1.43, 1.64, 3.85, 0.05, 0.82, 0.92, 2.01, 1.29)	Two-tailed p values per cluster = (2.73 x10^{-04}, 1.99 x10^{-01}, 1.57 x10^{-01}, 4.50 x10^{-04}, 9.62 x10^{-01}, 4.38 x10^{-01}, 4.05 x10^{-01}, 8.12 x10^{-02}, 2.36 x10^{-01})	Number of neurons per cluster = (160, 178, 95, 151, 147, 141, 176, 317, 280)	BH correction across clusters and conditions (i.e. tests were done for both sucrose and quinine)
Fig 4G	See figure	Pearson correlation	r = 0.766	Two-tailed p value = 0.0161	Number of clusters = 9	NA
Fig 4H	See figure	Pearson correlation	r = 0.709	Two-tailed p value = 0.0325	Number of clusters = 9	NA
Fig S1C	Difference in GLM coefficients between late and early in learning	Two sample t test (treated as independent here; see next row for analysis of only the longitudinally tracked neurons between these sessions)	t for all clusters = (-6.14, 0.10, -9.07, -1.88, 2.86, 1.25, -4.57, -6.46, -3.06)	Two-tailed p values per cluster = (8.14 x10^{-09}, 9.22 x10^{-01}, 8.64 x10^{-16}, 7.80 x10^{-02}, 6.71 x10^{-03}, 2.39 x10^{-01}, 1.66 x10^{-05}, 1.04 x10^{-09}, 4.52 x10^{-03})	Number of neurons per cluster Late = (476, 461, 499, 580, 422, 424, 606, 699, 646) Early = (202, 96, 137, 150, 190, 179, 201, 245, 190)	BH correction across clusters

Fig S1C	Difference in GLM coefficients between late and early in learning for longitudinally tracked neurons between these sessions	Paired t test	t for all clusters = (-7.43, 0.09, -7.54, -4.27, 1.27, 2.93, -5.11, -4.15, -2.81)	Two-tailed p values per cluster = (2.70×10^{-11}, 9.30×10^{-01}, 2.77×10^{-11}, 7.86×10^{-05}, 2.30×10^{-01}, 5.76×10^{-03}, 2.28×10^{-06}, 8.19×10^{-05}, 7.10×10^{-03})	Number of neurons per cluster = (202, 96, 137, 150, 190, 179, 201, 245, 190)	BH correction across clusters
Fig S1D	Difference in GLM coefficients between rewarded and unrewarded trials	Paired t test	t for all clusters = (-2.89, 4.81, -4.15, 0.72, 6.66, -2.06, -0.10, -0.52, -0.03)	Two-tailed p values per cluster = (9.11×10^{-03}, 9.23×10^{-06}, 1.17×10^{-04}, 7.09×10^{-01}, 8.02×10^{-10}, 7.25×10^{-02}, 9.76×10^{-01}, 7.74×10^{-01}, 9.76×10^{-01})	Number of neurons per cluster = (470, 458, 497, 572, 416, 423, 596, 693, 638)	BH correction across clusters
Fig S1E	Difference in GLM coefficients between late and early in learning	Two sample t test (treated as independent here; see next row for analysis of only the longitudinally tracked neurons between these sessions)	t for all clusters = (-2.78, -0.97, 0.10, -1.99, -1.44, -2.94)	Two-tailed p values per cluster = (2.62×10^{-02}, 4.06×10^{-01}, 9.24×10^{-01}, 9.85×10^{-02}, 2.31×10^{-01}, 2.58×10^{-02})	Number of neurons per cluster Late = (69, 79, 47, 91, 91, 89) Early = (32, 30, 28, 54, 45, 41)	BH correction across clusters
Fig S1E	Difference in GLM coefficients between late and early in learning for longitudinally tracked neurons between these sessions	Paired t test	t for all clusters = (-3.03, -1.48, 0.08, -2.07, -2.42, -1.71)	Two-tailed p values per cluster = ($3.03 \times 10^{--02}$, $1.80 \times 10^{--01}$, $9.38 \times 10^{--01}$, $8.72 \times 10^{--02}$, $5.97 \times 10^{--02}$, $1.43 \times 10^{--01}$)	Number of neurons per cluster = (30, 30, 27, 53, 45, 41)	BH correction across clusters

Fig S1F	Difference in GLM coefficients between rewarded and unrewarded trials	Paired t test	t for all clusters = (-0.45, 1.70, 3.24, 1.35, 0.37, 0.67)	Two-tailed p values per cluster = (7.12 x10^{-01}, 2.78 x10^{-01}, 1.35 x10^{-02}, 3.61 x10^{-01}, 7.12 x10^{-01}, 7.12 x10^{-01})	Number of neurons per cluster = (67, 78, 46, 89, 89, 87)	BH correction across clusters
Fig S2A	Mean fluorescence (0-3 s) for unpredicted reward delivered during intertrial interval	One sample t test	t for all clusters = (9.09, 13.87, 3.52, 4.81, 14.10, 10.32, 4.62, 5.89, 3.05)	Two-tailed p values per cluster = (9.40 x10^{-18}, 5.51 x10^{-34}, 5.28 x10^{-04}, 3.00 x10^{-06}, 1.06 x10^{-35}, 7.58 x10^{-22}, 6.22 x10^{-06}, 1.21 x10^{-08}, 2.41 x10^{-03})	Number of neurons per cluster = (414, 324, 456, 476, 383, 403, 511, 582, 517)	BH correction across clusters
Fig S2B	Difference in mean fluorescence (0-3 s) between sucrose with quinine and sucrose without quinine	Two sample t test (treated as independent here; see next row for analysis of only the longitudinally tracked neurons between these sessions)	t for all clusters = (-1.49, -0.78, 0.30, 0.17, 0.10, -1.43)	Two-tailed p values per cluster = (4.71 x10^{-01}, 8.80 x10^{-01}, 9.20 x10^{-01}, 9.20 x10^{-01}, 9.20 x10^{-01}, 4.71 x10^{-01})	Number of neurons per cluster sucrose with quinine = (57, 60, 40, 69, 74, 68) Sucrose without quinine = (32, 30, 28, 54, 45, 41)	BH correction across clusters
Fig S2B	Difference in mean fluorescence (0-3 s) between sucrose with quinine and sucrose without quinine for longitudinally tracked neurons between these sessions	Paired t test	t for all clusters = (-2.23, -2.37, 0.19, -0.99, -1.25, -1.85)	Two-tailed p values per cluster = (1.03 x10^{-01}, 1.03 x10^{-01}, 8.52 x10^{-01}, 3.92 x10^{-01}, 3.31 x10^{-01}, 1.47 x10^{-01})	Number of neurons per cluster = (27, 24, 24, 41, 35, 34)	BH correction across clusters

Fig S3F	Mean fluorescence of sucrose response (0-3 s) versus baseline (-2 to 0 s)	One sample t test	t(4) = 3.13	Two-tailed p value = 0.0351	Number of animals = 5	NA
Fig S4A	Mean difference in fluorescence response to sucrose with and without LED in a 2 s epoch	Paired t test	t for all clusters = (0.63, 0.63, 1.57, 1.13, 0.02, 0.56, 0.60, 1.16, 0.53)	Two-tailed p values per cluster = (6.73 x10^{-01}, 6.73 x10^{-01}, 6.73 x10^{-01}, 6.73 x10^{-01}, 9.87 x10^{-01}, 6.73 x10^{-01}, 6.73 x10^{-01}, 6.73 x10^{-01}, 6.73 x10^{-01})	Number of neurons per cluster = (82, 39, 48, 152, 52, 95, 227, 213, 281)	BH correction across clusters
Fig S4B	Mean difference in fluorescence response to spontaneous inhibition with respect to baseline before inhibition	One sample t test	t for all clusters = (1.52, 0.87, 2.36, 1.49, 2.48, 1.36, 1.54, -0.56, 0.99)	Two-tailed p values per cluster = (2.48 x10^{-01}, 4.36 x10^{-01}, 1.00 x10^{-01}, 2.48 x10^{-01}, 1.00 x10^{-01}, 2.64 x10^{-01}, 2.48 x10^{-01}, 5.79 x10^{-01}, 4.14 x10^{-01})	Number of neurons per cluster = (132, 74, 48, 112, 90, 88, 76, 189, 185)	BH correction across clusters
Fig S4D	Mean difference in fluorescence response to sucrose with and without LED in a 2 s epoch	Paired t test	t for all clusters = (1.38, 2.09, 1.40, 1.99, 2.72, 3.10, -1.86, 0.12, 1.77)	Two-tailed p values per cluster = (2.59 x10^{-01}, 1.08 x10^{-01}, 2.59 x10^{-01}, 1.08 x10^{-01}, 4.37 x10^{-02}, 4.37 x10^{-02}, 1.30 x10^{-01}, 9.54 x10^{-01}, 1.41 x10^{-01})	Number of neurons per cluster = (81, 39, 53, 153, 52, 96, 227, 216, 282)	BH correction across clusters and conditions (i.e. tests were done for both sucrose and quinine)
Fig S4D	Mean difference in fluorescence response to quinine with and without	Paired t test	t for all clusters = (0.01, 2.82, 0.97, 2.46, -0.61,	Two-tailed p values per cluster = (9.95 x10^{-01}, 4.37 x10^{-02}, 4.66 x10^{-01}, 4.50 x10^{-02},	Number of neurons per cluster = (81, 39, 53, 153, 52, 96, 227, 216, 282)	BH correction across clusters and conditions (i.e. tests were done for

	LED in a 2 s epoch		2.61, 0.61, 0.81, 2.52)	6.15 x10^{-01}, 4.37 x10^{-02}, 6.15 x10^{-01}, 5.35 x10^{-01}, 4.37 x10^{-02})		both sucrose and quinine)

Supplementary References

Kaifosh, P., Zaremba, J.D., Danielson, N.B., and Losonczy, A. (2014). SIMA: Python software for analysis of dynamic fluorescence imaging data. Front Neuroinform *8*, 80.

Otis, J.M., Zhu, M., Namboodiri, V.M.K., Cook, C.A., Kosyk, O., Matan, A.M., Ying, R., Hashikawa, Y., Hashikawa, K., Trujillo-Pisanty, I., et al. (2019). Paraventricular Thalamus Projection Neurons Integrate Cortical and Hypothalamic Signals for Cue-Reward Processing. Neuron *103*, 423-431.e4.

Resendez, S.L., Jennings, J.H., Ung, R.L., Namboodiri, V.M.K., Zhou, Z.C., Otis, J.M., Nomura, H., McHenry, J.A., Kosyk, O., and Stuber, G.D. (2016). Visualization of cortical, subcortical and deep brain neural circuit dynamics during naturalistic mammalian behavior with head-mounted microscopes and chronically implanted lenses. Nat Protoc *11*, 566–597.