Title: Neuronal activity in sensory cortex predicts the specificity of learning.

Authors: Katherine C. Wood[1], Christopher F. Angeloni[1,2], Karmi Oxman[1], Claudia Clopath[4], *Maria N. Geffen[1,2,3]

Affiliations: [1]Departments of Otorhinolaryngology: Head and Neck Surgery, University of Pennsylvania

[2]Department of Psychology, University of Pennsylvania

[3]Departments of Neurology and Neuroscience, University of Pennsylvania

[4]Department of Bioengineering, Imperial College London

*Lead contact and corresponding author:

Maria N. Geffen: mgeffen@pennmedicine.upenn.edu

Summary

Learning to avoid dangerous signals while preserving normal behavioral responses to safe stimuli is essential for everyday behavior and survival. Like other forms of learning, fear learning has a high level of inter-subject variability. Following an identical fear conditioning protocol, different subjects exhibit a range of fear specificity. Under high specificity, subjects specialize fear to only the paired (dangerous) stimulus, whereas under low specificity, subjects generalize fear to other (safe) sensory stimuli. Pathological fear generalization underlies emotional disorders, such as post-traumatic stress disorder. Despite decades of work, the neuronal basis that determines fear specificity level remains unknown. We identified the neuronal code that underlies variability in fear specificity. We performed longitudinal imaging of activity of neuronal ensembles in the auditory cortex of mice prior to and after the mice were subjected to differential fear conditioning. The neuronal code in the auditory cortex prior to learning predicted the level of specificity following fear learning across subjects. After fear learning, population neuronal responses were reorganized: the responses to the safe stimulus decreased, whereas the responses to the dangerous stimulus remained the same, rather than decreasing as in *pseudo*-conditioned subjects. The magnitude of these changes, however, did not correlate with learning specificity, suggesting that they did not reflect the fear memory. Together, our results identify a new, temporally restricted, function for cortical activity in associative learning. These results reconcile seemingly conflicting previous findings and provide for a neuronal code for determining individual patterns in learning.

Keywords: fear conditioning, auditory cortex, sensory systems, learning, computational model, imaging, sensory cortex, tuning curve, neurobiology, population coding.

Introduction

Learning allows our brain to adjust sensory representations based on environmental demands. Fear conditioning, in which a neutral stimulus is paired with an aversive stimulus, is a robust form of associative learning: exposure to just a few stimuli can lead to a fear response that lasts over the subject's lifetime (Johansen et al., 2011; Quirk et al., 1997). However, the same fear conditioning paradigm elicits different levels of learning specificity across subjects: whereas some subjects specialize the fear response to the paired stimulus, other subjects generalize fear across other, neutral, stimuli (Aizenberg and Geffen, 2013; Chapuis and Wilson, 2011; Li et al., 2008; Resnik et al., 2011). In pathological cases, the generalization of the fear response to stimuli in non-threatening situations can lead to conditions such as post-traumatic stress disorder (PTSD) (Jovanovic and Ressler, 2010; Mahan and Ressler, 2012) and anxiety (Krusemark and Li, 2012). Therefore, determining the neuronal basis for learning specificity following fear conditioning is important and can lead to improved understanding of the neuropathology of these disorders. Whereas much is known about how fear is associated with the paired stimulus, the neuronal mechanisms that determine the level of specificity of fear learning remain poorly understood. Our first goal was to determine the neuronal basis for the differential fear learning specificity across subjects.

Multiple studies suggest the auditory cortex (AC), a central structure in the auditory pathway, is involved in fear learning. Inactivation of AC chemically (Letzkus et al., 2011) or with optogenetics (Dalmay et al., 2019), as well as partial suppression of inhibition in AC (Aizenberg et al., 2015) during differential fear conditioning (DFC) led to decreased learning specificity. These observations suggest that AC may determine the level of learning specificity (Aizenberg et al., 2015), therefore we tested whether neuronal codes in AC *prior* to conditioning predict specificity of fear learning.

There is, however, considerable controversy around the role of AC *following* fear conditioning. Changes in stimulus representation in AC following association learning have been proposed to represent a fear memory trace, the strength of memory or the acuity of stimulus discrimination (Quirk et al., 1997; Weinberger, 2004; Weinberger and Diamond, 1987; Wigestrand et al., 2017). Whereas classical results suggested that the changes in stimulus representation amplified the difference between CS+ and CS-, more recent studies found little change in representation of CS+, coupled with reduction of responses to CS- and other tones (Edeline and Weinberger, 1993; Gillet et al., 2018; Kato et al., 2015; Ohl and Scheich, 1996). Furthermore, inactivation of the auditory cortex did not affect fear memory recall of tones (Aizenberg and Geffen, 2013; Dalmay et al., 2019), suggesting that AC may not be involved in fear memory retrieval. If AC were involved in fear memory retrieval, we would expect the amplitude of the plastic changes in sound representation to reflect the level of learning specificity across subjects. Therefore, our second goal was to test the role of the plastic changes in auditory cortex in shaping fear learning specificity across subjects.

To address these goals, we imaged the activity of neuronal ensembles in AC over weeks, prior to and following differential fear conditioning. First, we established the neuronal basis for differential learning specificity across subjects by finding that neuronal activity in AC prior to fear conditioning predicted the level of learning specificity. Second, we found that the plastic changes in AC following fear conditioning were not correlated with the level of learning specificity across subjects, suggesting that the role of AC in fear learning is restricted to the consolidation period and the plastic changes in AC do not represent fear memory. These findings refine our understanding of the neuronal code for variability in fear learning across subjects and reconcile seemingly conflicting previous results on the function of the auditory cortex in fear learning.

Results

Experimental setup

Whereas AC is involved in auditory fear conditioning, its role in the specificity of fear learning remains controversial. To establish the relationship between sound-evoked activity in the primary auditory cortex and differential fear conditioning (DFC), we recorded simultaneous neural activity from hundreds of neurons. We tracked the same neurons before and after DFC, using two-photon imaging of a virally expressed fluorescent calcium probe (GCaMP6 (Chen et al., 2013), Fig. 1). Longitudinal imaging of neuronal activity in large ensembles of neurons in AC before and after conditioning allowed us to compare the representation of the CS stimuli before and after learning (Fig. 2A).

We conditioned mice by exposure to a sequence of 10 repeats of two tones, one of which co-terminated with a foot-shock (CS+, 15 kHz), and one which did not (CS-, 11.4kHz). *Pseudo*-conditioned mice were presented with the same stimuli (CS, 11.4 kHz and 15 kHz), but the foot-shock occurred during periods of silence between the stimuli (Fig. 2B). We measured recall of the fear memory by presenting the same stimuli to the mouse in a different context and measuring the % of time the mouse froze (Fig. 2C). Learning specificity was defined as a difference between freezing to CS+ and CS- during memory recall tests (see Methods, Equation 1). Conditioned mice differentiated between the CS+ and CS- (H_0: learning specificity = 0, *t*-test, $t(13) = 2.53$, $p = .025$, Fig. 2D) with varying levels of specificity (range from -16.9 to 55.6%), whereas mice that were *pseudo*-conditioned showed little evidence of discriminating between the CS+ and CS- (H_0: learning specificity = 0, *t*-test, $t(10) = 0.83$, $p = .427$, range from -5.6 to 9.1%). Learning specificity persisted for the subsequent memory retrieval sessions over of the experiment (Fig.

2E, no significant effect of day after DFC1 on the learning specificity: $F_{(3,55)} = 0.08$, $p = .972$, conditioned mice). Thus, we found that mice exhibit different levels of learning specificity, with some generalizing and others specializing their fear responses to tones.

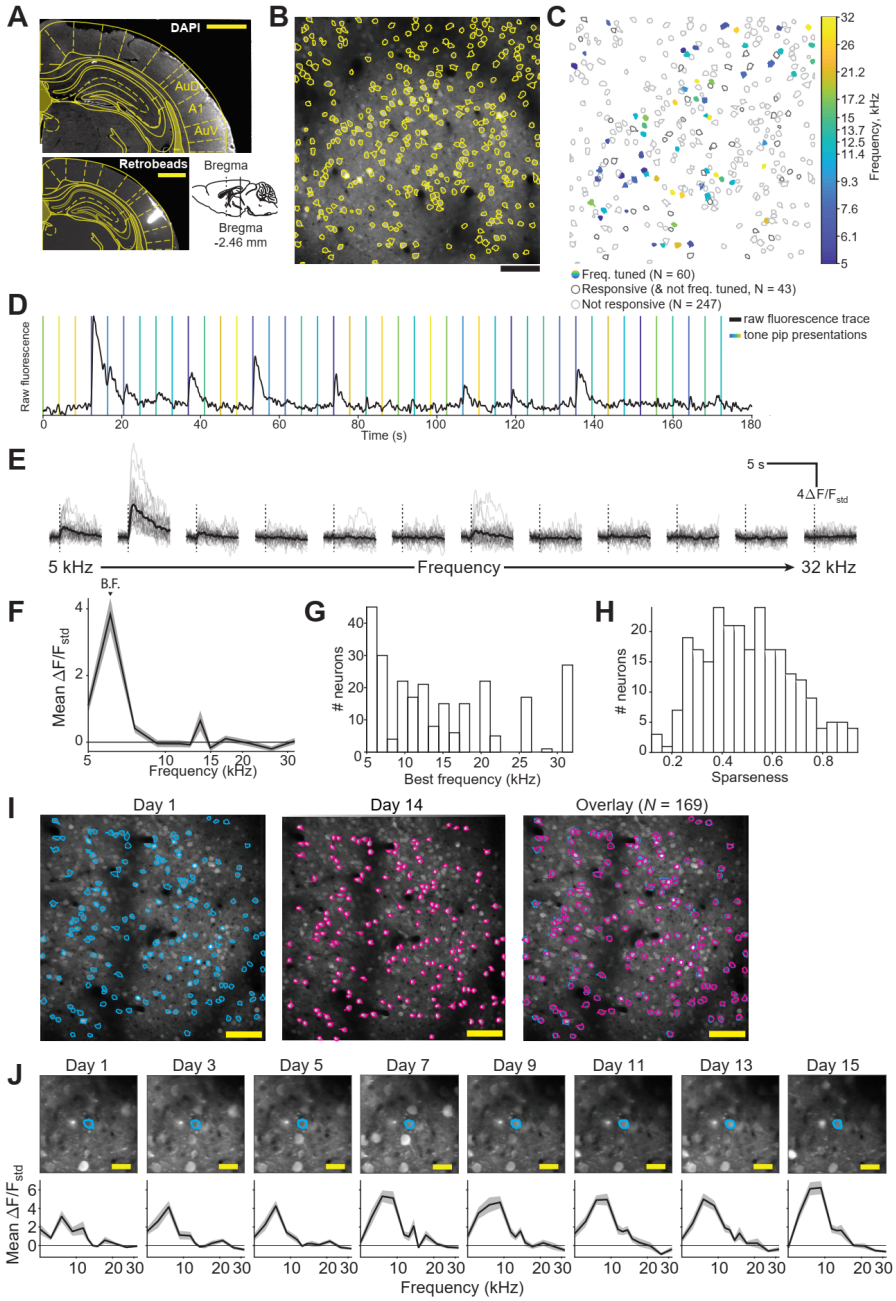

Figure 1: Longitudinal two-photon imaging tracks activity of neurons over weeks. (A) Example anatomical location of imaging, top panel: DAPI-stained (white, to show cell nuclei) section with mouse brain atlas overlaid (yellow). Bottom left panel: retrobeads (white) injection site into A1. The bottom right panel indicates the section's location relative to bregma in the sagittal plane. A1 = primary auditory cortex, AuD = dorsal auditory cortex, AuV = ventral auditory cortex. Scale bar = 1 mm. **(B)** Two-photon imaging field of view with regions of interest corresponding to individual neurons (yellow outline, N = 350). Scale bar = 100 µm. **(C)** Cell outlines from **B** indicating cells not responsive to the stimuli (light gray), cells responsive to tones (dark gray, t-test against zero, $p < .05$, corrected for multiple

comparisons) but not frequency tuned and frequency-tuned cells (colored according to frequency tuning, significantly responsive and one-way ANOVA, $p < .05$). Color bar indicates best frequency of each tuned neuron. **(D)** Part of a raw fluorescence trace (black) for an example frequency-tuned neuron with tone pip presentation times overlaid in color (vertical lines). The color of the vertical lines corresponds to the frequency of the tone pip presented – colors as in **C**. **(E)** Responses of neuron in **D** with single-trial responses (gray, $N = 25$ for each frequency) and the mean response (black). Dashed lines = tone pip onsets. **(F)** Mean response (from tone onset to 2 s after tone onset) across trials at each frequency of neuron in **D-E**. This neuron has a best frequency (B.F.) of 6.1 kHz. **(G)** Distribution of best frequencies of frequency-tuned cells recorded 24 hours pre-DFC ($N = 255$, mice = 25). **(H)** Distribution of sparseness of frequency-tuned cells recorded 24 hours pre-DFC. **(I)** Shows the field of view from two imaging sessions from the same mouse, 15 days apart (left and middle) with ROIs tracked between the two sessions outlined in cyan and magenta. The right panel shows the ROIs from the two sessions overlaid. Scale bar = 100 µm. **(J)** Frequency responses of a representative cell over the 8 sessions of the experiment. Cell is shown outlined (cyan line). Scale bar = 25 µm. Error bars = standard error of the mean (SEM).

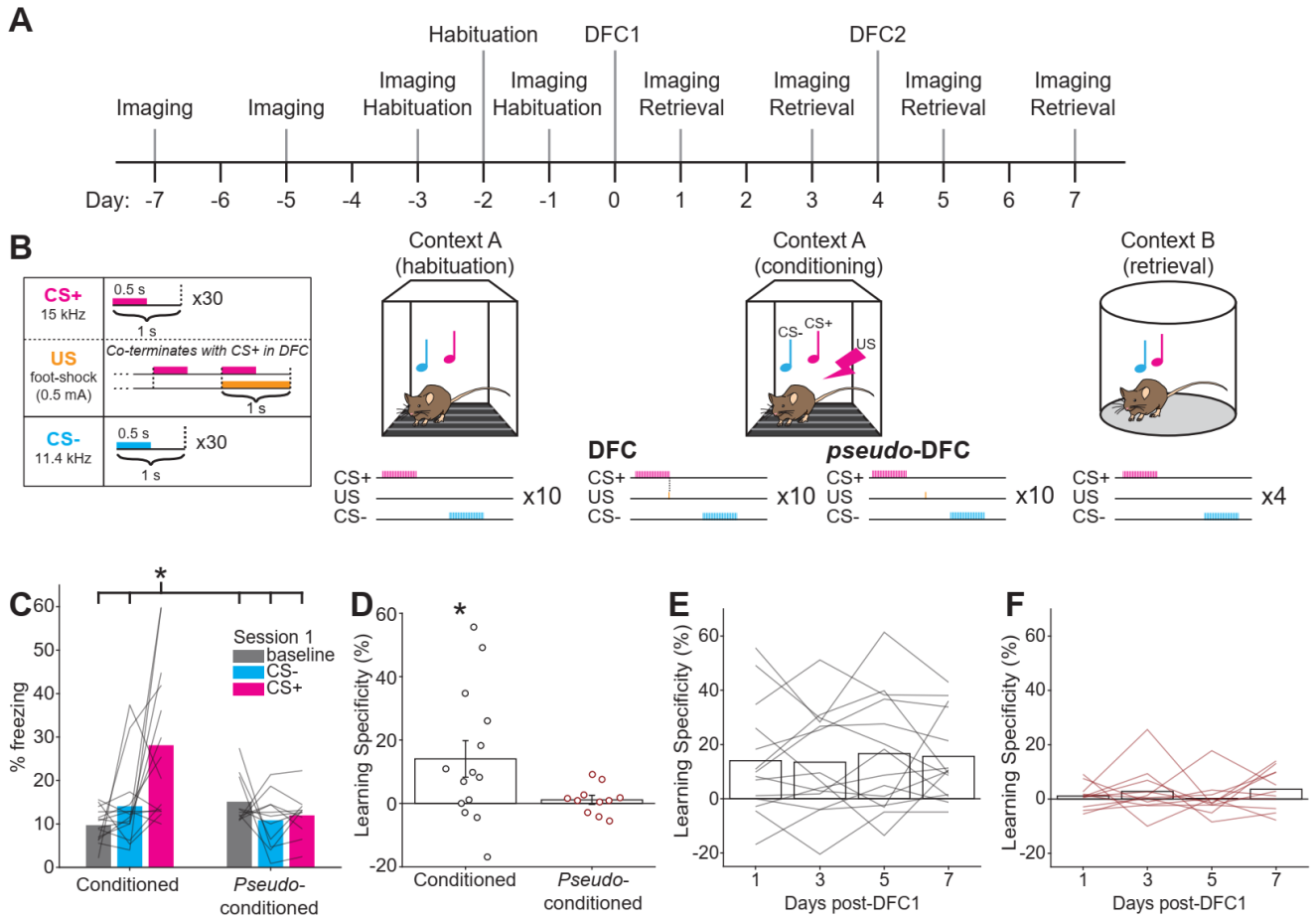

Figure 2: Experimental timeline and differential fear conditioning (DFC) paradigm. (A) Experimental timeline: Mice were imaged for 4 sessions (48 hours apart) prior to DFC to establish baseline responses to tone stimuli. Mice were subjected to DFC (14 mice) or *pseudo*-DFC (11 mice) on Days 0 and 4. After DFC1 (day 0), fear retrieval testing was performed after each imaging session. **(B)** During DFC (context A), a foot-shock (1 s, 0.3-0.5 mA) was paired with the CS+ (15 kHz, 30s pulsed at

1 Hz, 10 repeats). The CS- (11.4 kHz, 30 s pulsed at 10 Hz) was presented alternately with the CS+ (30-180 s apart, 10 repeats) and not paired with a foot-shock. During *pseudo*-DFC, 10 foot-shocks were presented randomly between the CS stimuli. During retrieval testing (context B), the same CS+ and CS- stimuli were presented alternately (30-180 s apart, 4 repeats), the motion of the mouse was recorded and the percentage freezing during each stimulus was measured offline. **(C)** Fear retrieval on day 1 (24 hours post-DFC1) showing the percentage of time frozen during tone presentation for CS+, CS- and baseline (mean 30 s prior to tone onset) for each mouse. (* Two-way ANOVA (Table S1), *Tukey-Kramer post-hoc* test, $p < .05$) **(D)** Mean learning specificity of conditioned and *pseudo*-conditioned mice. Circles show individual mice. (*t-test, $t(13) = 2.53$, $p = .025$) **(E)** Mean learning specificity for conditioned mice across days 1-7 post-DFC, black lines show individual mice. **(F)** Mean learning specificity for *pseudo*-conditioned mice across days 1-7 post-DFC, red lines show individual mice.

Neuronal responses to sounds in AC pre-DFC predict specificity of fear learning.

We hypothesized that activity in auditory cortex predicts differential learning specificity across individual mice. Specifically, we tested whether neuronal activity in AC pre-DFC1 predicted learning specificity post-DFC1. To assess how well single neurons could discriminate between the two conditioned tones, we computed the Z-score difference (Z_{diff}, see Equation 2 in Methods) of responses to CS+ and CS- for each neuron and each imaging session and calculated the mean Z_{diff} pre-DFC across neurons. In an example neuron (Fig. 3A), the distribution of single-trial response magnitudes to CS+ and CS- demonstrate a separation between the responses to the two CSs and resulting in a significant Z_{diff} score of 2.01. The Z_{diff} score of each cell was considered significant if the actual score was greater than the 95th percentile of the bootstrapped Z_{diff} scores (see Methods). To assess whether the neural discriminability could predict the learning specificity we calculated the correlation between the two measures, we did not find a significant correlation ($r(12) = .504$, 95% CI [0, 0.84], $p = .070$). However, when we averaged Z_{diff} scores across all 4 pre-DFC1 sessions (days -7 to -1) we found a strong correlation with learning specificity ($r(12) = .731$, 95% CI [.38, 0.93], $p = .011$). In summary, the neural discriminability of individual neurons in AC pre-DFC predicted the learning specificity 24 hours post-DFC.

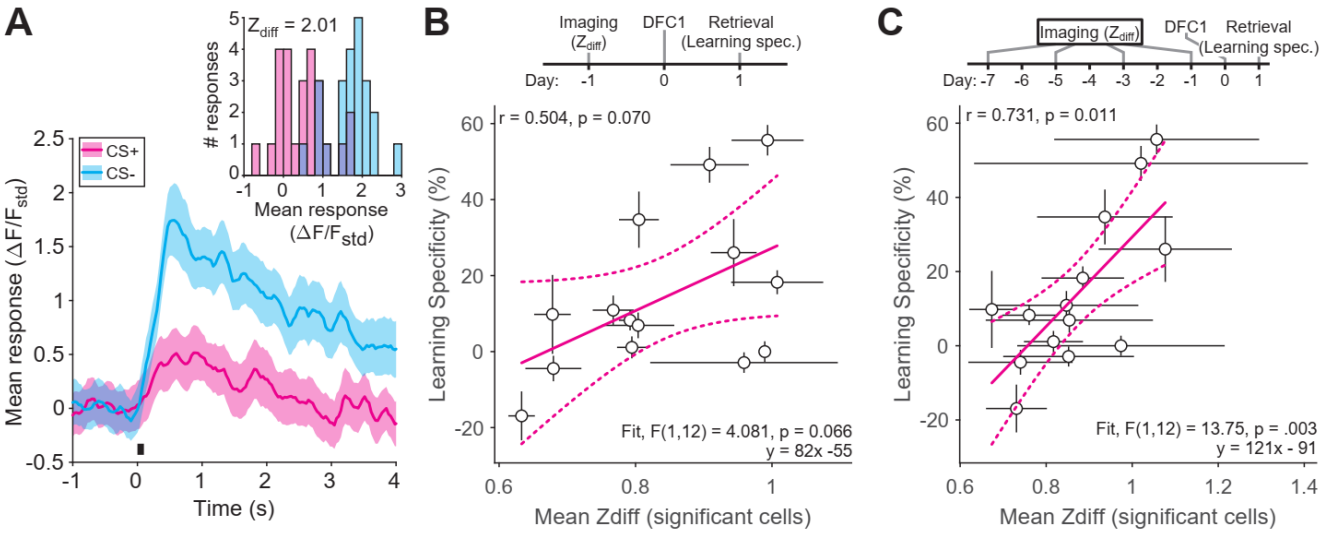

Figure 3. Z-scored difference (Z_{diff}) of responses to CS+ and CS- predicts learning specificity. (A) Response (mean + SEM, 25 repeats) to the presentation (black bar) of CS+ (magenta) and CS- (cyan) of

an example neuron. Inset shows distributions of the single-trial mean responses to CS+ and CS- from the same neuron. **(B)** Mean Z_{diff} score of neurons with significant Z_{diff} scores of each mouse 24 hours pre-DFC1 (day -1) predicts learning specificity 24 hours post-DFC1 (day 1). (**C**) Mean Z_{diff} score of significant neurons from the 4 pre-DFC (days -7 to -1) sessions predicts the learning specificity 24 hours post-DFC (day 1). Circles represent individual mice. In **B** and **C:** error bars = SEM. Magenta lines = linear regression (with 95% confidence intervals, dotted magenta lines) between the mean Z_{diff} and learning specificity. Pink text = *F*-test of linear regression *vs.* constant model with Z_{diff} score predicting learning specificity. Black text = Pearson's correlation test between Z_{diff} scores and learning specificity.

Population neuronal activity in AC prior to DFC predicts specificity of fear learning.

For many brain regions and tasks, activity of multiple neurons combined can provide more information in combination than averaged activity of individual neurons (Georgopoulos et al., 1986). We therefore investigated whether populations of neurons could predict learning specificity better than the average Z_{diff} scores using machine learning. We trained a Support Vector Machine (SVM, Fig. 4A), using a linear kernel, to discriminate between presentation of CS+ and CS- with 10-fold cross-validation using population responses to the CS+ and CS- from the imaging session 24 hours pre-DFC1 (day -1). We found a significant correlation between the SVM performance and the learning specificity from 24 hours post-DFC1 (Fig. 4B, $r(12) = .540$, 95% CI [.23, .82], $p = .046$). Thus, learning specificity could be predicted based on the SVM performance just 24 hours pre-DFC1. When averaging the mean SVM performance across the 4 sessions preceding DFC1 (days -7 to -1; Fig. 4C), we also observed a significant correlation ($r(12) = .752$, 95% CI [.36, .92], $p = .003$). This correlation coefficient was not significantly stronger than that of day -1 alone (r difference = -0.212, 95% CI [-.35, -.062], $p = .108$). Thus, a linear measure of population activity that takes into account all neurons could predict learning specificity from a single, as well as multiple sessions.

Neuronal information may be integrated in a non-linear fashion in the brain across neurons (Stringer et al., 2019). Therefore, we next tested whether a non-linear combination of information in populations of neurons provides additional information and thus predicts learning specificity better than a linear SVM. We found that there was no significant difference (bootstrap test, see methods, r difference = .000, 95% CI [-.34, .19], $p = .964$) between the correlation coefficients of learning specificity with SVM performance using either a linear kernel (Fig. 4C) or a non-linear kernel (Gaussian, $r(12) = .751$, 95% CI [.49, .92], $p = .003$). Thus, using a non-linear kernel with the SVM did not improve prediction of the learning specificity. Furthermore, consistent with the notion that cortical computation is thought to largely rely on linear integration of inputs (DiCarlo et al., 2012), we found that the mean Z_{diff} score across individual neurons of each mouse strongly correlated with the SVM decoding performance (Fig. 4D, Pearson's correlation; $r(12) = .842$, 95% CI [.70, .94], $p = .001$). These combined results suggest that activity of neuronal population in AC predicts learning specificity by combining information from neurons in a linear fashion.

It is possible that neurons that do not provide information about the two CS stimuli are only contributing noise to the SVM model. Thus, we hypothesized that if neurons with non-significant Z_{diff} scores were merely adding noise to the SVM model, then using only neurons with significant Z_{diff} scores in the model would lead to increased performance and better prediction of learning specificity than using all neurons. Thus, we trained the SVM with only the cells that had significant Z_{diff} scores and found improved SVM prediction performance compared with using all responsive cells (Fig. 4E, paired *t*-test, $t(13) = -8.1$, $p < .001$). However, there was no significant difference in correlation (r difference = .037,

95% CI [-.40, .13], p = .702) between the learning specificity and performance of the SVM using all neurons (Fig. 4C) or only neurons with significant Z_{diff} scores (Fig. 4F, $r(12)$ = .789, 95% CI [.55, .92], p = .001). This analysis suggests that a linear combination of responses of all, rather than select AC neuronal responses, likely determines the level of learning specificity.

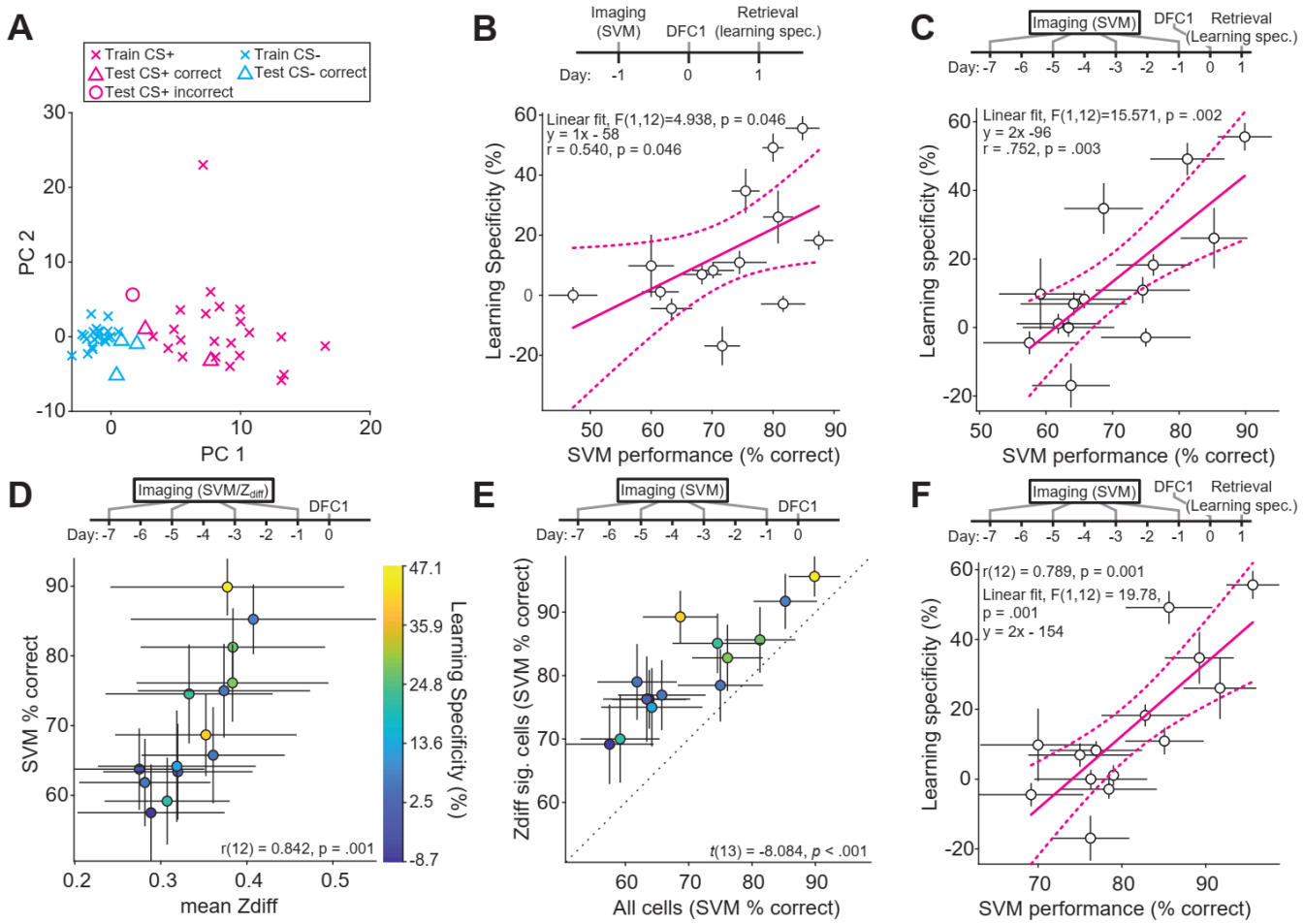

Figure 4. Neural population discrimination between CS+ and CS- pre-DFC predicts learning specificity. (A) Example training (crosses), correct test (triangles) and incorrect test (circles) data for the SVM projected onto the first two principal components of the population responses to each trial. Training and testing data consisted of population responses (mean response for each cell) to CS- (cyan) and CS+ (magenta). **(B)** SVM performance 24 hours pre-DFC1 (day -1) predicts learning specificity after DFC1 (Day 1, N = 14). **(C)** Mean SVM performance across the 4 sessions pre-DFC1 (days -7 to -1) predicts learning specificity post-DFC1 (day 1). **(D)** Performance of the SVM correlates with the mean Z_{diff}. Statistics: Pearson correlation. Fill color indicates learning specificity. **(E)** Performance of the SVM across sessions pre-DFC1 (days -7 to -1) increases when using only cells with significant Z_{diff}. Statistics: paired t-test between mean Z_{diff} and SVM performance. Dashed line = line of unity. **(F)** SVM performance across sessions pre-DFC1 (days -7 to -1) using only cells with significant Z_{diff} scores predicts learning specificity. Magenta lines in **B**, **C**, **D** & **F** show the linear fit between the two variables (95% CI, dashed magenta lines). Error bars show standard error of the mean.

Neuronal activity after DFC does not predict fear retrieval.

We next tested the temporal extent for the prediction of learning specificity based on the neuronal responses in AC from each imaging session after the first conditioning session (DFC1) and the following memory retrieval session. If changes in sound-evoked responses in AC reflect memory formation or the strength of learning as previously suggested (Edeline and Weinberger, 1993), we expect a strengthened relationship between the neural discrimination and learning specificity after DFC than prior to DFC. However, we found no significant correlation between the significant Z_{diff} scores in each session after DFC1 (days 1, 3 & 5) and the learning specificity from the following sessions (days 3, 5 & 7, Fig. S2A). The mean significant Z_{diff} score no longer predicted the learning specificity after DFC. We found similar results with the population neural discrimination; there were no significant correlations between the SVM performances in sessions after DFC1 (days 1, 3 & 5) and the learning specificity from the following sessions (days 3, 5 & 7, Fig. S2D-F). Thus, SVM performance did not predict learning specificity following DFC.

Taken together with the evidence that two measures of neural discriminability pre-DFC can predict subsequent learning specificity, these data show that cortical neural discriminability pre-DFC, but not post-DFC, determined learning specificity. This suggests that changes in neuronal activity in AC do not necessarily reflect the formation of a fear memory or the strength of the memory. We next sought to understand how stimulus representation changes in AC after DFC.

Suppression of responses in AC to non-behaviorally-relevant stimuli post-DFC.

We tested whether AC exhibited changes in responses to CS+ and CS- and whether these changes were dependent on learning. After conditioning, select neurons in AC amplified the difference between CS+ and CS-, similar to the previous reports of select neurons (Edeline and Weinberger, 1993). However, across the population, there were heterogeneous changes in neuronal responses (Fig. 5A, Table S2). On average, the responses of neurons to CS+ were preserved, whereas the response at CS- decreased (Fig. 5B, *t*-test; $p < .05$ (corrected for multiple comparisons using false discovery rate– see methods), Table S3). In contrast, in *pseudo*-conditioned mice, mean responses at most frequencies, including both CS frequencies, decreased (Fig. 5C, Table S4), while individual neurons exhibited heterogenous changes (Fig. 5A). These results contradict the theory that following fear conditioning, the reorganization of neuronal activity serves to amplify the difference in responses to CS+ and CS- thereby improving discriminability (Edeline and Weinberger, 1993). Rather than increasing the responses at CS+ and decreasing the responses at CS-, the plasticity serves to counteract reduction in activity that occurs in the absence of foot-shock pairing at the paired tone frequency.

To identify the relationship between the changes in AC responses at CS+ and CS- and fear learning, we tested whether these changes were caused by changes in positive or negative fluorescence responses, with negative responses indicating inhibition of the response. In cells significantly responding pre- or post-DFC1 split by positive and negative responses (Fig. 5D), for conditioned mice, the decrease in response at CS- was mainly driven by a decrease in positive responses. In *pseudo*-conditioned mice, the decrease in response at both CS stimuli was driven by decreases in positive responses (Fig. 5E). Overall, these results suggest that responses to non-behaviorally relevant stimuli are reduced overall (CS- in conditioned mice and both CS in *pseudo*-conditoned mice), whereas responses to the paired and therefore behaviorally relevant stimuli (CS+ in conditioned mice) are maintained.

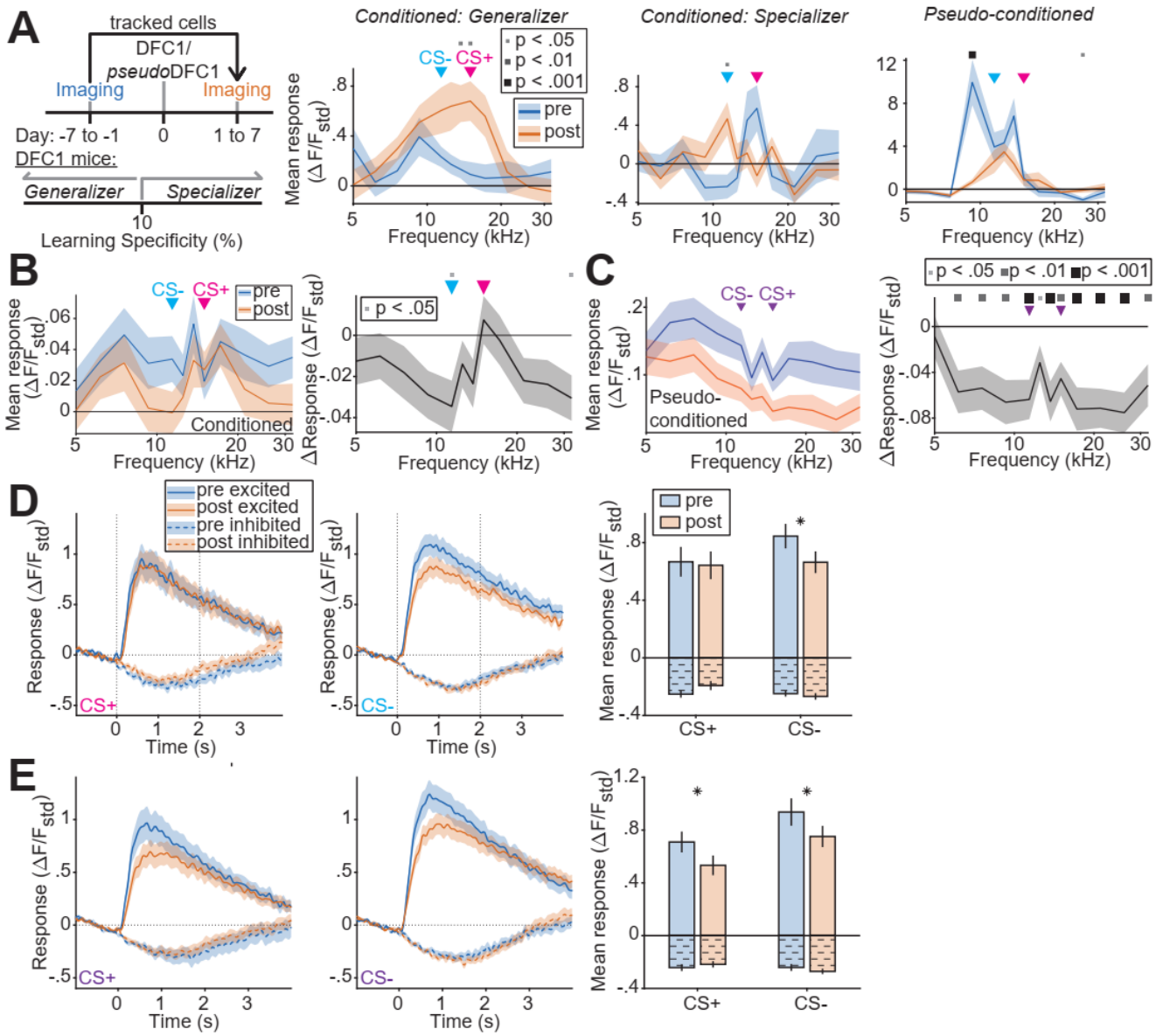

Figure 5: Changes in frequency representation post-DFC1. (A) We tracked the responses of the same cells pre- (day -1) and post-DFC1 (day 1). Conditioned mice with a learning specificity >10 were classified as specializers whereas if learning specificity was <10, they were classified as generalizers. The right panels show three example frequency tuning curves from tracked neurons from conditioned and *pseudo*-conditioned mice 24 hours pre-DFC (day -1, blue) and 24 hours post-DFC (day 1, orange). Significant differences in the tuning curves are indicated by the square sizes above (paired *t*-test, corrected for multiple comparisons – see methods). The two arrows show the frequencies of the CS- (11.4 kHz) and CS+ (15 kHz). **(B)** (Left panel) Mean frequency tuning curves across all conditioned mice ($N = 14$) of cells tracked and responsive on at least one session pre- and post-DFC1 ($N = 1013$). (right panel) Mean change in reponse at each frequency of the same cells, squares indicate significant changes (paired *t*-test, corrected for multiple comparisons). **(C)** (Left panel) Mean frequency tuning curves across all *pseudo*-conditioned mice ($N = 11$) and cells tracked and responsive on at least one session pre- and post-*pseudo*-DFC1 ($N = 729$). (right panel) Mean change in response pre- and post-*pseudo*-DFC1 at each frequency for the same cells, square sizes indicate significant changes (paired *t*-test, corrected for multiple comparisons). **(D)**

Mean responses of cells from conditioned mice presented with CS+ and CS- in the frequency-tuning stimuli ($N = 10$) that significantly responded to CS+ (left panel) and CS- (middle panel) split by whether the response was positive (solid lines) or negative (dashed lines) 24 hours pre- (blue) and post-DFC1 (orange). Dashed vertical lines indicate the window responses were averaged over. (right panel) Mean responses (within a 2-s window after sound onset). **(E)** Mean responses of cells from *pseudo*-conditioned mice ($N = 11$) that significantly responded to CS+ (left panel) and CS- (middle panel) split by whether the response was positive (solid lines) or negative (dashed lines) 24 hours pre- (blue) and post-DFC1 (orange). Dashed vertical lines indicate the window responses were averaged over. (right panel) Mean responses (within a 2-s window after sound onset).

Previous studies found that the best frequency of neurons shifts towards the conditioned stimulus (CS+) after DFC (Edeline and Weinberger, 1993). In conditioned mice, we did not observe consistent changes in best frequency towards the CS+. On average there was no change in the best frequency of frequency-tuned neurons from pre- to post-DFC1 (Fig. 6A, paired *t*-test, $t(648) = -0.166$, $p = .868$). Sparseness decreasd from pre- to post-DFC1 suggesting an increase in tuning width (Fig. 6B, paired *t*-test, $t(648) = -2.868$, $p = .004$). Similarly, in *pseudo*-conditioned mice, the best frequency did not change (Fig. 6C, paired *t*-test, $t(523) = 0.917$, $p = .360$), but sparseness decreased (Fig. 6D, paired *t*-test, $t(523) = -3.136$, $p = .002$). Overall, the best frequency of neurons was unaffected by DFC in both conditioned and *pseudo*-conditioned mice and the sparseness decreased in both sets of mice. suggesting there were no changes specific to DFC in frequency tuning properties.

In summary, we observed heterogeneous changes in response of individual cells tracked from pre- to post-DFC1. In conditioned animals, there was on average a decrease in mean response at CS- and no change at the CS+. In *pseudo*-conditioned mice, we observed decreases in response at the CS- and CS+. These findings are consistent with the theory that responses to non-behaviorally relevant stimuli are suppressed, possibly due to long-term habituation (Gillet et al., 2018; Kato et al., 2015). Despite these changes, neurons did not exhibit a shift in the best frequency toward the CS+, as may have been expected from previous work (Bakin and Weinberger, 1990; Edeline and Weinberger, 1993).

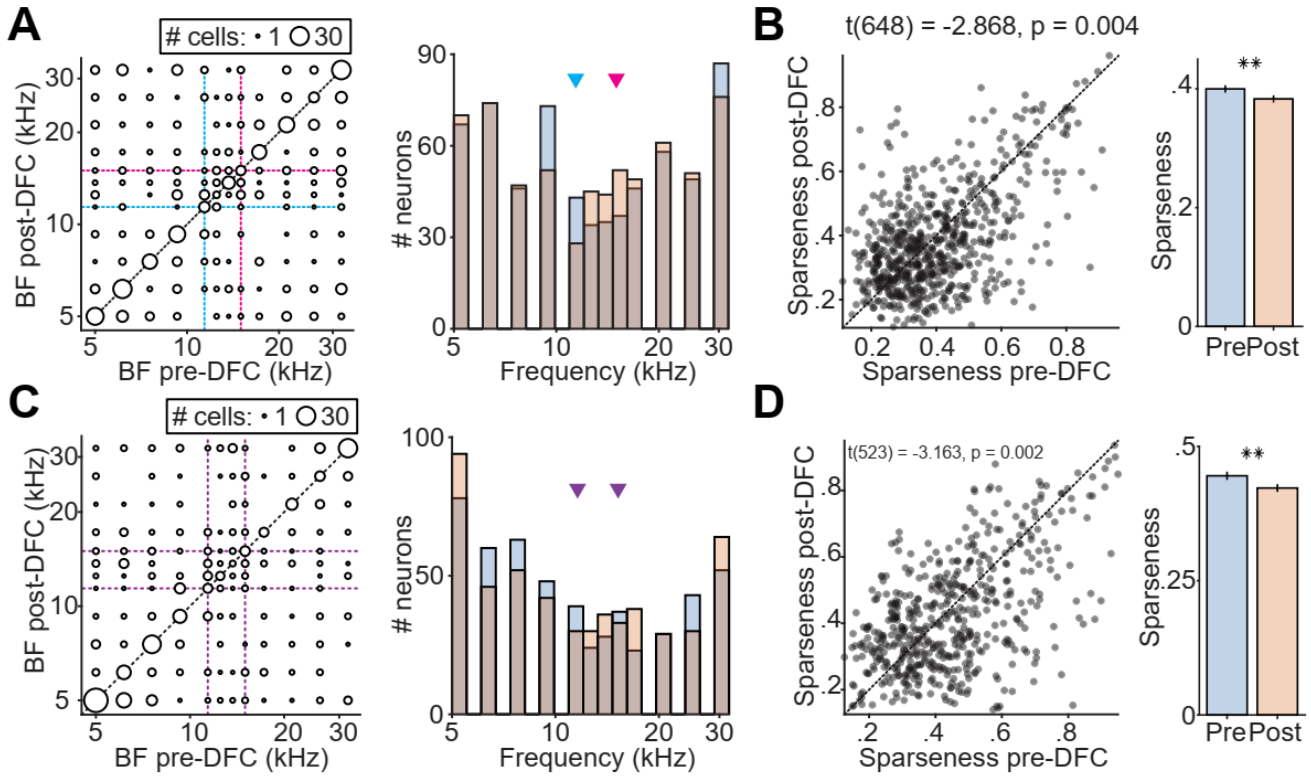

Figure 6: Best frequency and sparseness pre- and post-conditioning. (A) (left panel) Best frequencies pre- and post-DFC1. (right panel) Distribution of best frequencies pre- and post-DFC1. **(B)** (left panel) Sparseness pre- and post-DFC1. Circles = individual cells. (right panel) Mean sparseness pre- and post-DFC1. **(C)** (left panel) Best frequencies pre- and post-*pseudo*-DFC1. (right panel) Distribution of best frequencies pre- and post-*pseudo*-DFC1. **(D)** (left panel) Sparseness pre- and post-*pseudo*-DFC1. (right panel) Mean sparseness pre- and post-*pseudo*-DFC1.

Fear conditioning leads to preservation of discriminability between CS+ and CS-

It is possible that the lack of change in response at CS+ and the decrease in response at CS- could lead to improved discriminability between CS+ and CS- in conditioned mice by increasing the difference between the responses to each stimulus. Therefore, we tested whether the neuronal discriminability of CS+ and CS- was improved after DFC1. We found that, on average, there was no change in the mean Z_{diff} score from pre- to post-DFC1 (Fig. 7A. Paired *t*-test, $t(13) = 0.774$, $p = .453$) suggesting that neural discriminability between CS+ and CS- did not improve. Furthermore, we found no change in the SVM performance from pre- to post-DFC1 (Fig. 7B. Paired *t*-test, $t(13) = 0.849$, $p = .411$), suggesting that, at the neuronal population level, there was also no improvement in discriminability between CS+ and CS-. Thus, although responses to CS- decreased and responses to CS+ remained constant, this change did not result in greater discriminability at population level.

To further investigate whether the changes in stimulus representation were related to the DFC, we tested the performance of the SVM on each imaging session using cells tracked across consecutive imaging sessions. We trained the SVM using the first session (*n*) and tested on data held out from that

session and from the following session (n + 1). If there were a significant reorganization of stimulus representation between days that contributes to stimulus discrimination, we would expect a change in performance following DFC1. We did not observe any consistent change in SVM performance following DFC1 (Fig. 7C, paired t-test, $t(12) = 1.43$, $p = .176$) nor did we observe any significant changes over any of the consecutive sessions of the experiment (Fig. 7D, see Table S5 for statistical results). Similarly, on average there was no change in Z_{diff} scores in cells tracked between consecutive sessions (Fig. S3, Table S6). Overall, these results indicate that changes that occurred in stimulus representation do not contribute to the neuronal stimulus discrimination in fear memory retrieval.

Different levels of learning specificity across mice could potentially account for the different levels of neuronal discriminability post-DFC. We therefore tested whether there was a correlation between the neural discriminability (mean Z_{diff} score and SVM performance) and the learning specificity *post-DFC*. The mean Z_{diff} score did not correlate with the mean learning specificity of each mouse across the 4 post-DFC1 sessions (Fig. 7E, Pearson correlation, $r(12) = .130$, CI [-.32, .63], $p = .675$), nor was there a correlation between the mean SVM performance post-DFC and the mean learning specificity post-DFC (Fig. 7F, $r(12) = .106$, CI [-.44, .74] $p = .713$). This suggests that neuronal discriminability post-DFC does not reflect learning specificity.

It is conceivable that the *pseudo*-conditioned mice learned that both the CS+ and CS- stimuli were not behaviorally relevant, or that they were both 'safe' sounds, since a foot-shock could not occur during their presentation. If responses to behaviorally irrelevant stimuli are suppressed (perhaps by habituation Gillet et al., 2018; Kato et al., 2015) then the neural discrimination of the two CS stimuli would be impaired after DFC. Indeed, we found that the mean Z_{diff} scores decreased (Fig. 7G, paired t-test, $t(10) = -2.274$, $p = .046$). However, there was no change in SVM performance after *pseudo*-DFC in *pseudo*-conditioned mice (Fig. 7H, paired t-test, $t(10) = -1.932$, $p = .082$). Combined, these results suggest that neural discriminability in fear-conditioned mice is preserved, counteracting reduction in responses observed in *pseudo*-conditioned mice.

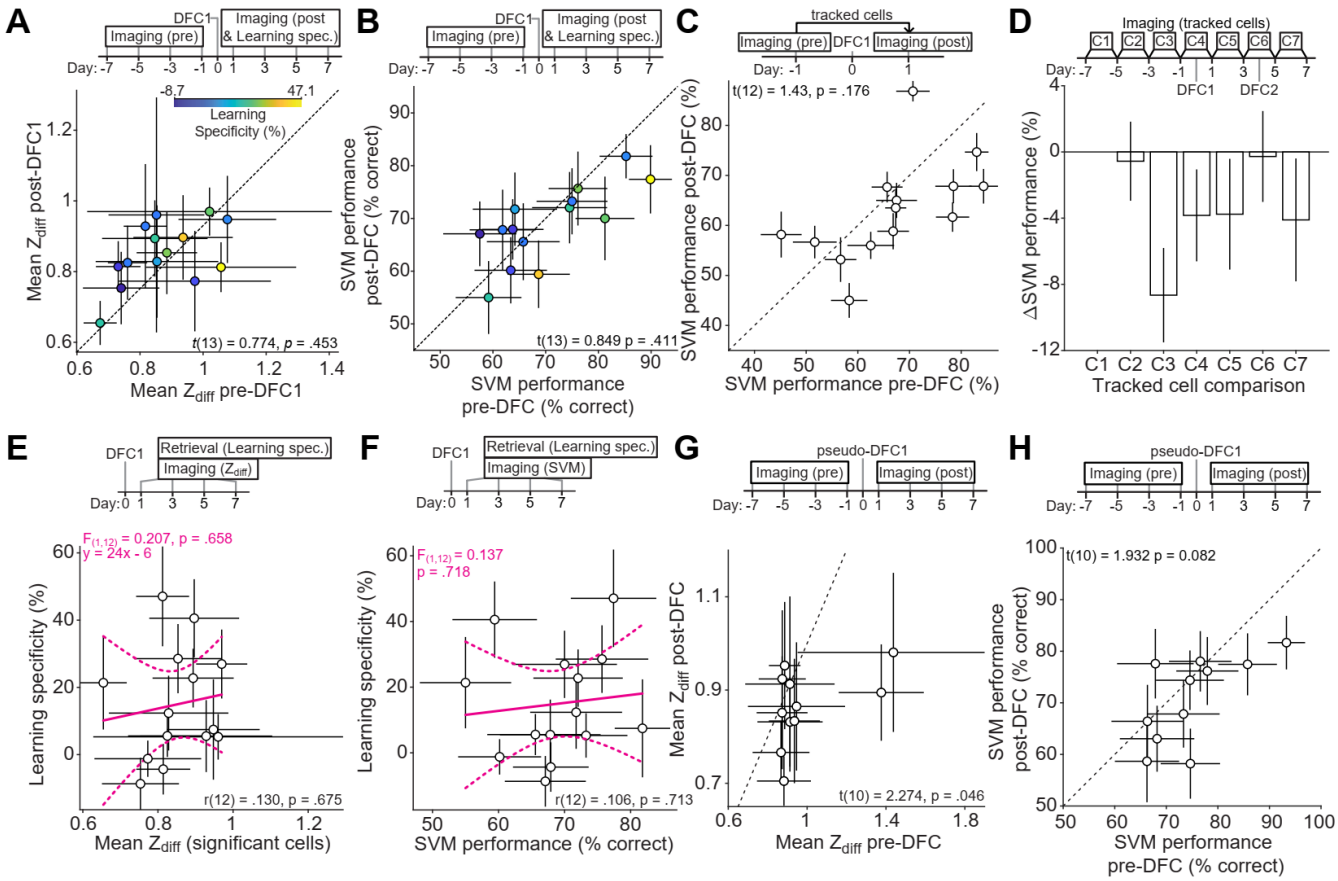

Figure 7: Changes in stimulus information representation post-DAFC. (A) Comparison of Z_{diff} between the pre- (days -7 to -1) and post-DFC1 sessions (days 1 to 7) in conditioned mice. Color represents average learning specificity (days 1-7). Statistics: paired *t*-test. **(B)** Comparison of SVM performance between the pre- (days -7 to -1) and post-DFC1 sessions (days 1 to 7) in conditioned mice. Color represents learning specificity as in **A.** Statistics: paired *t*-test. **(C)** SVM performance 24 hours pre- (day -1) and post-DFC1 (day 1) trained using data from day -1 and tested on data from both days. **(D)** Mean change in SVM performance using tracked cells between all consecutive imaging sessions of the experiment as in **C**. For example, comparison 4 (C4, mean change in SVM performance of data from **C**) is the mean of SVM performance on day 1 – SVM performance on day -1. **(E)** Across the 4 post-DFC1 sessions (days 1-7), the mean Z_{diff} score does not correlate with the mean learning specificity in conditioned mice. Magenta line indicates linear fit (dashed magenta = 95% CI). **(F)** Across the 4 post-DFC1 sessions (days 1-7), the mean SVM performance does not correlate with learning specificity in conditioned mice. Magenta line indicates linear fit (dashed magenta = 95% CI). **(G)** Comparing mean Z_{diff} score across the pre- (days -7 to -1) and post-DFC1 sessions (days 1-7) in *pseudo*-conditioned mice. Statistics: paired *t*-test. **(H)** Comparing mean SVM performance across the pre- (days -7 to -1) and post-DFC1 sessions (days 1-7) in *pseudo*-conditioned mice. Statistics: paired *t*-test . In all panels, circles show each mouse, all error bars are SEM

A learning model of the fear circuit

We found that AC activity prior to learning predicts specificity of learning, yet the reorganization of neural responses does not correlate with learning specificity. This could seem contradictory to previous studies showing that decreasing inhibition in AC decreased learning specificity (Aizenberg et al., 2015); but that following conditioning, silencing AC did not affect learning specificity (Aizenberg and Geffen, 2013; Dalmay et al., 2019). How can AC predict learning specificity, but at the same time how can AC silencing not affect learning specificy? In order to better understand our findings in relation with previous results, we built a simple model that consisted of two frequency-tuned populations of neurons and a neural population that responds to the foot-shock. Our goal was to test whether this simple model could account for both the findings in this manuscript and the results from previous work, in particular: (1) Discriminability between CS+ and CS- in AC predicts learning specificity post-DFC (Fig. 3-4). (2) Suppressing inhibition in AC leads to increased generalization (decreased learning specificity) post-DFC (Aizenberg et al., 2015). (3) Suppressing AC post-DFC does not affect learning specificity (Aizenberg and Geffen, 2013; Dalmay et al., 2019).

In the model, we included two populations of frequency-tuned neurons (representing the medial geniculate body, MGB, and auditory cortex, AC). The MGB receives auditory inputs and projects to AC. Both populations project to basolateral amygdala (BLA). AC sends tonotopically organized feedback connections to MGB. During conditioning, MGB neurons receive sound inputs and the neurons in the BLA are active during the foot-shock (Fig. 7A). The weights from MGB and AC to BLA are updated according to a Delta learning rule (see Methods); that is, they are potentiated when both are co-activated (i.e. when the foot-shock coincides with the sound stimulus). We control the level of overlap in frequency tuning between neurons in AC and use it as a parameter representing frequency discriminability (more overlap = less discriminability). The activity of the BLA after weight update and with auditory input only is then used as a measure of freezing.

We first tested whether broad tuning in AC, which results in lower discriminability between CS+ and CS-, produced more generalized freezing than sharp tuning. Indeed, we found that increased overlap in frequency tuning in AC neurons, without changing the tuning of MGB neurons, drove more generalized freezing responses (Fig. 8B, Figure S4). This is due to the fact that, when AC was broadly tuned, CS+ tone activated AC neurons not only responding to CS+ frequency but also to other frequencies, such as CS-, albeit to a lesser extent. After learning, this resulted in strong AC to BLA synaptic weights that are not specific to CS+. MGB is narrowly tuned in our model, but the weights from MGB to BLA were also strengthed in a non-specific fashion (Fig. S4) because AC projects back to MGB. Therefore, CS+ also activated non-specific neurons in MGB concurrently with the foot-shock. These results support the present findings (Fig. 3-4). Second, we examined the effects of decreasing inhibition in the AC population during conditioning (Fig. 8C, Fig. S5). Decreasing inhibition resulted in an increased overlap in frequency responses in the AC population, which in turn led to increased generalization, supporting and providing a mechanism for previous findings (Aizenberg et al., 2015; Briguglio et al., 2018). Third, we tested whether inactivating the auditory cortex following conditioning had an effect on freezing responses (Fig. 7C, Fig. S6). Consistent with previous findings (Aizenberg and Geffen, 2013; Dalmay et al., 2019), we did not observe a change in fear generalization following AC inactivation. The broad or narrow tuning of AC neurons allowed for the synapses from MGB to BLA to be strengthened either narrowly or broadly during conditioning. Therefore, with suppression of AC after conditioning during memory recall, the specialized versus generalized learning was preserved.

Combined, the model demonstrates that a simple anatomically consistent circuit supports multiple aspects of cortical control of fear conditioning previously identified. Discriminability between CS+ and CS- in AC predicts learning specificity and influences learning in the MGB to BLA synapses. After learning, AC is not necessary for fear memory retrieval.

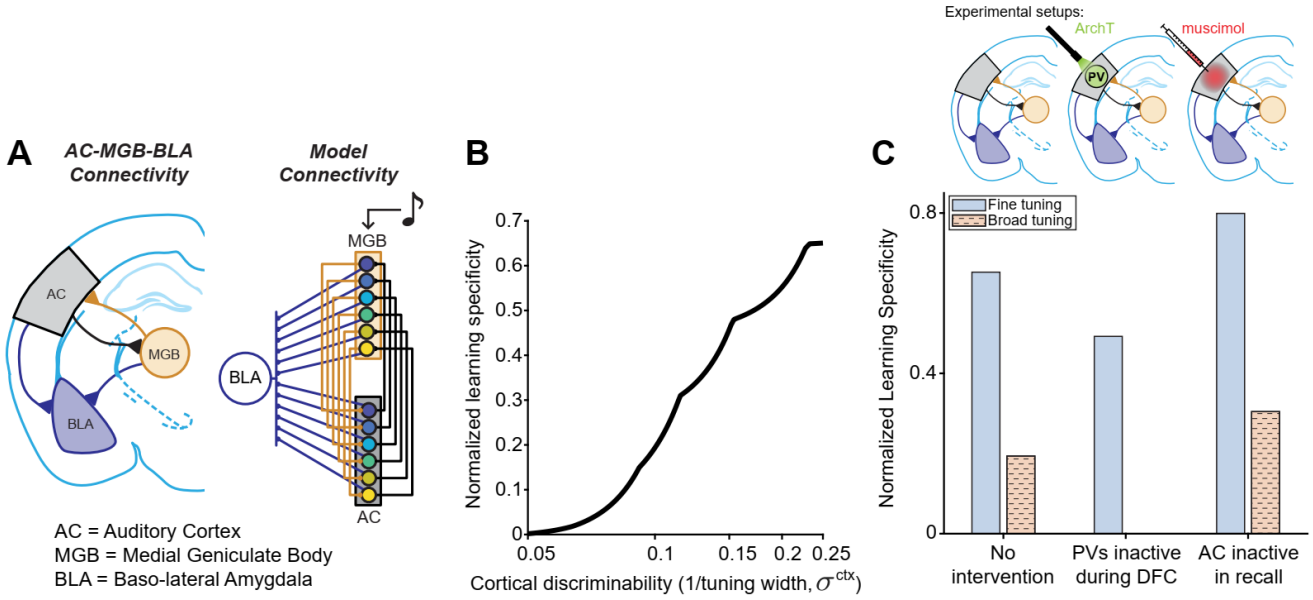

Figure 8: A learning model reconciles present and past findings. (A) (left panel) connectivity between auditory cortex (AC, gray), medial geniculate body (MGB, orange) and basolateral amygdala (BLA, blue). Right panel shows the model connectivity. MGB provides input to AC (orange lines), and both MGB and AC provide inputs to BLA (blue lines). AC feeds back to MGB (black lines). Colored circles represent neurons tuned to different, overlapping frequency ranges. **(B)** Normalized learning specificity output from the model with varying levels of AC discriminability, achieved by changing the frequency tuning overlap (σ^{ctx}) between the neurons in the AC population. **(C)** Normalized learning specificity at two AC discriminability levels; fine (blue) and broad tuning (orange). Results are shown for learning specificity with no interventions, when inhibition is reduced in AC during DFC (analogue of when ArchT-transfected PV interneurons in AC are inactivated by optogenetics during DFC), and when AC is inactivated during memory recall (analogue of an injection of muscimol during memory recall; PV = parvalbumin positive interneurons, ArchT = Archaerhodopsin-T).

Discussion

Learning to fear dangerous sensory signals while maintaining normal behavioral responses to safe or neutral signals is important for everyday behavior and survival. Abnormalities in fear learning can lead to disorders such as PTSD, in which fear is generalized to neutral stimuli. Even in normal subjects, fear learning can lead to fear responses that are specialized to specific stimuli or generalized across neutral stimuli. Whereas much progress has been made on identifying the neuronal mechanisms for fear learning, little is understood about the neuronal mechanisms that determine learning specificity. Understanding the neuronal mechanisms that control specialization and generalization in fear learning is essential and may lead to improved understanding of the neural pathology of fear disorders.

Previous work suggests that the sensory cortex plays a role in discrimination of simple tone stimuli during fear memory acquisition (Aizenberg et al., 2015; Briguglio et al., 2018; Wigestrand et al., 2017). In this study, we identified the role of the auditory cortex in discriminative fear learning. We tracked the representation of a dangerous and a safe stimulus in the auditory cortex of mice before and after differential fear conditioning. As expected, mice exhibited a range of specificity of the fear response, with some exhibiting fear to only the dangerous tone, and others exhibiting fear to both stimuli (Fig. 2). Importantly, we found that the neural discriminability of the safe and dangerous sounds in the auditory cortex before DFC predicted the subsequent learning specificity (Fig. 3, 4). Furthermore, we identified changes in the neuronal representation of the sounds in AC after fear conditioning; responses to safe stimulus in conditioned mice and both unpaired tones in *pseudo*-conditioned mice were suppressed, whereas responses to the dangerous tone were unchanged in conditioned mice (Fig. 5, 6). However, these changes in neuronal representation did not correlate with the specificity of learning, suggesting that the changes in the auditory cortex do not reflect the fear memory as previously suggested (Fig. 7). Combined, our results demonstrate that: (1) Prior to fear learning, neuronal responses in AC shape fear specificity; (2) Following fear conditioning, neuronal response transformations are not correlated with fear learning specificity, and therefore the auditory cortex does not encode auditory differential fear memory; (3) The neuronal response transformations are consistent with maintenance of responses to behaviorally relevant stimuli that counteract habituation to non-relevant stimuli; (4) A simple model of the auditory nuclei and the basolateral amygdala (BLA) could account for our results as well as a number of previous findings (Fig. 8). Our study reconciled multiple conflicting results from prior studies identifying the role and limitations of the involvement of the sensory cortex in differential fear learning. The results provide evidence for a complex function of sensory cortex in learning and memory.

Our finding that the neuronal activity prior to fear conditioning predicted specialization of fear learning provides a mechanism for a number of previous observations. Previous work suggested a role of AC in differential fear memory acquisition (Aizenberg et al., 2015; Briguglio et al., 2018; Letzkus et al., 2011; Wigestrand et al., 2017). Specifically, inactivation of inhibitory neurons in the auditory cortex during fear conditioning led to an increase in generalization of fear learning (Aizenberg et al., 2015). Suppressing inhibitory neurons in the auditory cortex led to a decrease in Fisher information which reflects the certainty about a stimulus in neuronal representation (Briguglio et al., 2018). This change would likely result in a decrease in neuronal discriminability between the CS+ and the CS- tones in the auditory cortex, and therefore drive a decrease in fear specificity, as demonstrated by our model (Fig. 8). Our results provide the link between optogenetic inactivation of interneurons in AC leading to increased fear generalization, and to increased frequency tuning width, which decreases neuronal discriminability.

Our finding that changes in neuronal representation of the stimuli after fear conditioning did not correlate with learning specificity puts in perspective multiple previous studies and suggests that the auditory cortex does not reflect the differential fear memory. Previous work found that changes in neuronal responses to the dangerous and safe stimuli after differential fear conditioning amplified the difference between the responses (Edeline and Weinberger, 1993; Ohl and Scheich, 1996). This change was proposed to represent fear memory (Ohl and Scheich, 1996, 2004; Weinberger, 2004). We identified similar transformations in a subset of neurons that were tracked pre- to post-conditioning. However, we found that over the neuronal population, these changes abolished the correlation between neuronal discrimination and learning specificity. This observation suggests that neuronal code in the auditory cortex after fear conditioning does not reflect differential fear memory. Indeed, a number of studies found that inactivating the auditory cortex after fear conditioning does not affect fear memory retrieval (Aizenberg

and Geffen, 2013; Dalmay et al., 2019), but see (Gillet et al., 2018; Wigestrand et al., 2017). Combined, our results restrict the role of auditory cortex in fear conditioning to differential fear memory acquisition, but not retrieval.

The neuronal changes following differential fear conditioning can be interpreted as a relief of habituation in response to the dangerous tones, and habituation for the safe sounds. Over the neuronal population, we found that neurons responsive to the dangerous stimulus maintained their responses, while neurons responsive to the safe stimulus reduced their responses after differential fear conditioning. By contrast, in a control *pseudo*-conditioned group of mice that were exposed to the sound and aversive stimuli in an unpaired fashion, the responses to both stimuli were reduced. These findings are consistent with a previous study (Gillet et al., 2018), which proposed that the reduction in response to safe stimuli (not behaviorally relevant) is due to habituation of neuronal responses to repeated, non-behaviorally relevant stimuli (Kato et al., 2017) and maintenance of responses to dangerous stimuli is thus a counteraction of habituation. Despite the apparent increase in neural contrast between the CS+ and CS- in conditioned mice, neural discrimination measures did not change from pre- to post-DFC. However, we did observe a decrease in neural discrimination measures post-DFC in *pseudo*-conditioned mice. This suggests that neural activity in the cortex was reorganized so as to maintain discrimination of the behaviorally relevant stimuli while maintaining frequency discrimination across other frequencies, thus supporting stable auditory perception.

We found that a simple model with connections from auditory nuclei to the basolateral amygdala could replicate the results found in the current study and account for previous work (Fig. 8). The model demonstrated that (1) neural activity in cortex can predict subsequent learning specificity; that (2) inactivation of PV interneurons in AC during DFC leads to increased generalization (Aizenberg et al., 2015), and that (3) the auditory cortex is not necessary for differential fear memory retrieval (Aizenberg and Geffen, 2013; Dalmay et al., 2019). The model proposes that either MGB or AC or a combination of both can induce auditory fear memory through the strengthening of connections in the amygdala. We propose that feedback from auditory cortex to the MGB contributes to discrimination of perceptually similar stimuli during DFC by controlling stimulus discrimination in the MGB, this may or may not be a direct projection neuroanatomically (He, 2003; Suga, 2008).

In the model, an auditory nucleus, MGB, was reciprocally connected with AC in the circuit, as reported experimentally (Chen et al., 2019b; Williamson and Polley, 2019). The MGB is also strongly implicated in fear conditioning (Antunes and Moita, 2010; Apergis-Schoute et al., 2005; Herry and Johansen, 2014; although, see Suga, 2008; Weinberger, 2011) and is involved in consolidation of fear memory (Taylor et al., 2020). And, MGB does project extensively to auditory cortical areas and to the amygdala (Ledoux, 2000), where its projections converge with those from A1 (Lee, 2015) and from temporal association areas (Edeline, 1999), where AC and MGB may provide complementary information (Chen et al., 2019a). Furthermore, our recent results identified a pathway from the BLA to the MGB via the thalamic reticular nucleus, a thin sheet of inhibitory neurons, which may facilitate the amplification of sound-evoked responses or change in tuning width in the MGB (Aizenberg et al., 2019). Future studies need to explore the role of the MGB and specific projections between AC and MGB and BLA in fear learning and memory.

This study had several methodological limitations. Our results relied on tracking the neuronal responses in transfected neurons in AC and did not distinguish between different neuronal subtypes. We used a viral vector that expressed the fluorescent calcium indicator in all neurons, we were thus unable to

identify neuronal cell types in our recordings. Previous studies found that a specific class of inhibitory neurons increases activity with presentation of repeated tones (Gillet et al., 2018). It is therefore plausible that our results include a subset of neurons that function differentially during fear conditioning, but which we are unable to identify due to limited sample and lack of selective labelling. The role of specific neuronal subtypes in differential fear conditioning needs to be explored further in studies that target viral expression to these subsets. Furthermore, we restricted our recordings to superficial layers 2 and 3 of the auditory cortex, and it is possible our results overlook more specific changes in the thalamo-recipient layers of the cortex (Atencio et al., 2012; Linden and Schreiner, 2003). The complexity of transformations in the cortical microcircuit with learning should be explored further (Blackwell and Geffen, 2017; Harris and Shepherd, 2015; Wood et al., 2017).

The results of the study may be restricted to the pure tone stimuli, which are relatively simple auditory stimuli. Other studies have questioned the role of auditory cortex in classical fear conditioning, where the CS+ paired with foot-shock is presented in isolation (Armony et al., 1997; Romanski and LeDoux, 1992; Zhang et al., 2018); and identified differential function for the auditory cortex depending on the stimulus complexity in fear learning (Dalmay et al., 2019; Letzkus et al., 2011). It is likely that such an important behavioral modification as fear has redundant pathways to obtain the same outcome (Betley et al., 2013; Boatman and Kim, 2006; Zhang et al., 2018). The role of auditory cortex in discrimination of simple and complex stimuli has recently been a subject of interest in the field (Ceballo et al., 2019; O'Sullivan et al., 2019). Future studies will need to dissect to what extent the differences in neuronal codes in AC shape differential fear learning of complex and natural sounds.

Our results may be applicable to understanding anxiety disorders. An extreme example of fear generalization is realized in PTSD (Dunsmoor and Paz, 2015). Here we find that the present state of each individual brain, in terms of neural discrimination of stimuli, is predictive of the future generalization of fear in the subject. This suggests that a way to prevent generalization of dangerous and safe sounds is to improve neural discrimination of potentially threatening stimuli (Ginat-Frolich et al., 2017; Lange et al., 2017; Roesmann et al., 2020; Tuominen et al., 2019). Further work in this area can lead to a deeper understanding how genetic and social factors, as well early life experiences, shape cortical activity in this common and devastating disorder (Mahan and Ressler, 2012; Roesmann et al., 2020).

We found that the mammalian sensory cortex plays key role in stimulus discrimination during, but not following, differential fear conditioning. These results reconcile several previous findings and suggest that the role of sensory cortex is more complex than previously thought. Investigating the changes in the cortico-amygdalar circuit during fear learning will pave way for new findings on the mechanisms of learning and memory.

Acknowledgements

The authors thank Dr. Yale Cohen, Dr. Steve Eliades and Dr. Jay Gottfried on helpful discussions and comments on an earlier version of the manuscript. The authors also thank the members of the Geffen laboratory for their helpful advice. This work was supported by funding from the National Institute of Health grants R01DC015527, R01DC014479 and R01NS113241 to MNG.

Author Contributions

Conceptualization, M.N.G. and K.C.W.; Methodology, K.C.W, M.N.G., C.F.A. and C.C.; Software, K.C.W.; Investigation, K.C.W. and K.O..; Formal Analysis, K.C.W. and M.N.G; Writing – Original Draft, K.C.W. and M.N.G.; Writing – Review and Editing, K.C.W., M.N.G. and C.C.; Funding Acquisition, M.N.G.; Resources, M.N.G., K.O. and K.C.W.; Supervision, M.N.G.

Declaration of Interests

The authors declare no competing interests.

STAR methods

Mice

All experimental procedures were in accordance with NIH guidelines and approved by the IACUC at the University of Pennsylvania. Mice were acquired from Jackson Laboratories (16 male, 9 female, mean age of cranial window implant: 9.6 weeks [6.3 – 13.0 weeks]; PV-Cre (4) [Stock No: 017320], PV-Cre x ROSA (1) [Stock No: 003474], CamKII-Cre mice (1) [Stock No: 005359] or Cdh-23 mice (19) [Stock No: 018399]) and were housed in a room with a reversed light cycle. Experiments were carried out during the dark period. Mice were housed individually after the cranial window implant. 14 mice (9 male, 5 female) were in the conditioning group and 11 mice (7 male, 4 female) were in the *pseudo*-conditioning control group.

The Auditory Brainstem Response (ABR) to broadband clicks (2 – 80 kHz, 70 dB SPL) was acquired at the end of the experiment when possible (average 18.7 days post-final imaging session) in order to confirm that the mice had good hearing. The click ABR amplitude did not change significantly over the course of the experiments (Figure S7, mean ± s.d. pre = 1.84 ± 0.54 μV, post = 1.37 ± 0.34 μV, $N = 8$, paired t-test, $t(14) = 2.11$ $p = .053$).

Euthanasia procedures were consistent with the recommendations of the American Veterinary Medical Association (AVMA) Guidelines on Euthanasia.

Surgical procedures

Mice were implanted with cranial windows over auditory cortex. Briefly, the mice were anaesthetized with 1.5 – 3% isoflurane and a 3-mm circular craniotomy was performed over the left auditory cortex (stereotaxic coordinates) using a 3-mm biopsy punch centered over the stereotaxic coordinates of A1 (70% of the distance between bregma and lambda, 4.3 mm lateral to the midline). An adeno-associated virus (AAV) vector encoding the calcium indicator GCaMP6s or GCaMP6m (AAV1.Syn.GCaMP6s.WPRE.SV40 or AAV1.Syn.GCaMP6m.WPRE.SV40, UPENN vector core) was injected (750 nl, 1.89 x 10^{-12} genome copies·ml^{-1}) at a 750μm depth from the surface of the brain at 60 nl·min^{-1} for expression in layer 2/3 neurons in A1. 3 injections were made at the same lateral distance but separated by 0.5 mm in the anterior-posterior direction or 5 injections were made spread across the window (0.3 – 0.5 mm apart). The injection needle was left in place for 10 mins after the injection was complete before retraction. Injections were made using pulled (P-97 Puller, Sutter Instruments, USA) glass pipettes (Harvard Apparatus, USA) with tip openings of 30 – 50 μm using a pump (Pump 11 Elite, Harvard Apparatus, USA). After injection a circular 3-mm diameter glass coverslip (size 0 or 1, Warner Instruments) was placed in the craniotomy and fixed in place using a mix of cyanoacrylate glue and dental

cement. A custom-made stainless-steel head-plate (eMachine Shop) was fixed to the skull using C&B Metabond dental cement (Parkell). The implant was further secured using black dental cement. Mice were allowed to recover for 3 days post-surgery.

Behavioral training and testing

Mice underwent a minimum of 4 imaging sessions (range: 4 – 11) prior to differential auditory fear conditioning (DFC). DFC and subsequent fear retrieval testing took place in two different contexts (A and B, discussed below). Before and after each conditioning or retrieval, we cleaned the conditioning and testing chambers with either detergent (retrieval chamber) or 70% ethanol (conditioning chamber). We recorded a video of the mouse in the testing chamber using FreezeFrame 3 software (Coulbourn) at 3.75 Hz; the subsequent movement index (mean grayscale values of frame (n+1) minus the preceding frame (n)) was exported and analyzed offline using MATLAB. The threshold of movement was defined as the 12.5th percentile of the values from each session. The mouse was considered to be freezing if the movement index was below the threshold; the measure of freezing was expressed as a percentage of time spent freezing during stimulus presentation and for baseline during the 30s prior to stimulus onset.

Stimuli were generated using FreezeFrame 3 and presented at 70 dB SPL from an electrostatic speaker (ES-1, TDT) mounted above the animal. DFC took place in context A. Stimuli were 30 s in duration and were either a continuous pure tone (4 mice) or pulsed pure tones (500 ms duration at 1 Hz). The CS+ (15 kHz) was paired with a foot-shock (1 s, direct current, 0.5 – 0.7 mA, 10 pairings, inter-trial interval: 50 – 200 s) delivered through the floor of context A (by precision animal shocker, Coulbourn). The foot-shock either co-terminated with the continuous tone or the onset coincided with the final tone pulse of the CS+ stimuli. The CS- (11.4 kHz) was presented after each CS+-foot-shock pairing but was not reinforced (10 presentations, inter-trial interval: 20 – 180 s). The next day, after a two-photon imaging session, conditioned mice were tested for fear retrieval in context B, during which 4 presentations of the CS+ and CS- were presented (30 s duration, interleaved, inter-trial interval: 30 – 180 s). For 4 mice, longer continuous presentations of the CS+ and CS- were presented (either 120 s, 1 mouse, or 60 s, 3 mice), for these mice, trials were divided into 4 equal durations and treated as above. In *pseudo*-conditioning, the foot-shocks were presented interleaved between the stimuli in periods of silence.

For each mouse the learning specificity (*LS*, Equation 1 (Aizenberg and Geffen, 2013)) was calculated as:

$$LS = \sum_{i=1}^{N} fr_{CS^+}(i) \Big/ N - \sum_{i=1}^{N} fr_{CS^-}(i) \Big/ N$$

Equation 1

Where i is the trial index, $fr_{CS^{+/-}}(i)$ is the fraction of time spent freezing during trial i in the CS+/- condition, respectively, and N is the number of trials per condition.

Calcium imaging procedure and acoustic stimuli

All imaging sessions were carried out inside a single-walled acoustic isolation booth (Industrial Acoustics). Mice were placed in the imaging setup, and the head plate was secured to a custom base (eMachine Shop) serving to immobilize the head. Mice were gradually habituated to head-fixing over 3 – 5 days, 3 – 4 weeks after surgery and before imaging commenced. Imaging took place in mice aged 17.5 – 19.6 weeks (min: 12.9, max: 27.1 weeks).

We recorded changes in fluorescence of GCaMP6s/m caused by fluctuations in calcium concentration in transfected neurons of awake, head-fixed mice, using two-photon microscopy (Ultima *in vivo* multiphoton microscope, Bruker). The laser (940 nm, Chameleon Ti-Sapphire) power at the brain surface was kept below 30 mW. Recordings were made at 512 x 512 pixels and 13-bit resolution at ~30 frames per second.

Stimuli were generated at a sampling rate of 400 kHz using MATLAB (MathWorks, USA) and consisted of 100-ms long tone pips in the 5−32-kHz frequency range and presented at 60 – 80 dB SPL. In a single recording session, each frequency was repeated 15 – 25 times in a *pseudo*-random order with a 4-s inter-stimulus interval.

Data analysis and statistical procedures

Publicly available toolboxes (Pachitariu et al., 2017) running on MATLAB were used to register the two-photon images, select regions of interest (ROI) and estimate neuropil contamination, resulting in a neuropil-corrected fluorescence trace (F) for each neuron. We calculated the mean baseline fluorescence ($F_{baseline}$) and standard deviation of the baseline (F_{std}) over the one second prior to tone onset for each sound presentation, and then determined the change in fluorescence over time relative to the mean baseline fluorescence ($\Delta F = F - F_{baseline}$). We then divided ΔF by F_{std}, effectively calculating the z-score of the fluorescence relative to the baseline ($\Delta F/F_{std}$) for each sound presentation.

We imaged the activity from the same cells over 15 days in layer 2/3 of auditory cortex, using blood vessel architecture, depth from the surface and the shape of cells to return to the same imaging site. To identify ROIs across imaging sessions that corresponded to the same cell, the maximum-projection fluorescence images from each day were registered by transforming the coordinates of landmarks present in both images in MATLAB (2017a) using the *fitgeotrans* function. The transformation was applied to ROIs from the second imaging session to match the first – all subsequent sessions were aligned to the first imaging session. We next calculated the distance between all the pairs of centroids (mean x-y position of each ROI) across the two sessions; ROIs from the two sessions were then automatically registered as the same cell based on the nearest centroid. We then manually checked the shape and position of the ROIs for any pairs that had duplicate matches, <80% ROI overlap or a larger than average distance between the centroid locations. ROIs which were not matched to any earlier ROIs were counted as new cells and assigned a new cell number. This process was repeated for subsequent sessions, registering the imaging field to the first session and comparing the ROIs to the cumulative ROIs from previous sessions. A final manual inspection of all the unique ROIs was performed after all the imaging sessions were registered. ROIs that overlapped with each other extensively were excluded from the dataset since it was unclear whether they were the same or different cells. Examples of tracked cells and aligned ROIs are shown in Figure 1.

The response to each tone was defined as the mean $\Delta F/F_{std}$ over 2 seconds following tone onset. Neurons were deemed sound responsive if at least one of the frequency responses was different from zero (*t*-test, $p < 0.05$, corrected for multiple comparisons using false discovery rate (Benjamini and Hochberg, 1995; Groppe, 2020)) The frequency tuning curve was defined as the mean response to each tone frequency across repeats. Neurons were defined as frequency tuned if they were sound responsive and their tone responses were significantly modulated by tone frequency (one-way ANOVA, $p < .05$). Best frequency was defined as the frequency with the highest mean response. Sparseness (*S*, Equation 2 (Rust

and DiCarlo, 2012)) was used to estimate the sharpness of tuning curves, with 1 being very sharply tuned while 0 would indicate an equal response to each tone frequency:

$$a = \frac{((\sum r_i)/N)^2}{\sum(r_i^2/N)} \qquad S = \frac{1-a}{1-1/N}$$

Equation 2

Where r_i is the mean response to the frequency i and N is the total number of frequencies tested.

The Z-scored difference between responses to CS+ and CS- (Z_{diff}, Equation 3) was calculated for each neuron using the following equation:

$$Z_{diff} = \left| \frac{\sum r_{CS^+}/N - \sum r_{CS^-}/N}{\sqrt{\left(\sigma_{r_{CS^+}} \cdot \sigma_{r_{CS^-}}\right)}} \right|$$

Equation 3

Where $r_{CS+/CS-}$ is the single trial mean responses to CS+ and CS- respectively, N is the number of repeats of each stimulus and σ is the standard deviation of mean responses. The Z_{diff} score was considered significant if the actual Z_{diff} was larger than the 95th percentile of the distribution of Z_{diff} scores calculated from resampling the data 250 times with replacement while shuffling the CS+/CS- label of each response.

For fitting the Support Vector Machine, we used MATLAB's *fitcsvm* function with a linear kernel to predict the learning specificity based on the single-trial population responses (mean $\Delta F/F_{std}$ over 2 s post-stimulus onset for each neuron). We used 10-fold cross validation in testing performance of the model. For non-linear model testing, we used a Gaussian kernel, with a kernel scale of 18 to train the SVM.

We calculated significance of correlations using a bootstrap procedure, resampling the data 10000 times and computing the Pearson's correlation between the resampled data. We defined the 95% confidence limits of the correlation coefficient (r) as the 2.5th and 97.5th percentiles of the resulting distribution of correlation coefficients. In order to assess whether two correlations were significantly different from one another we subtracted the r distributions of each dataset from one another, the change in r was considered significant if 95% CI of the distribution did not overlap with zero.

To test whether the learning specificity in conditioned mice changed over time we performed an analysis of variance (ANOVA, using MATLAB) with learning specificity as the dependent variable and number of days after DFC1 (days 1, 3, 5 and 7) as the independent variable. Changes in frequency tuning were assessed using paired t-tests at each frequency between the mean pre- and post- response of each cell (t-test, $p < 0.05$, corrected for multiple comparisons using false discovery rate (Groppe, 2020)). For mice that were not tested at 11.4 and 15 kHz under the two-photon microscope (4 conditioned mice) responses were interpolated from the frequency tuning curves pre- and post-DFC. For cells present in more than one session either pre- or post-DFC, mean responses at each frequency were combined and the changes in response were assessed from the mean across pre- and post-DFC sessions. For comparing the fluorescence responses, the 4 mice not tested directly were excluded.

Confirming anatomical location of recording

Upon conclusion of the imaging sessions, we removed the windows of the mice and injected a red fluorescent marker (Red Retrobeads, CTB or AAV5.CAG.hChR2(H134R)-mCherry.WPRE.SV40

(mCherry)) into the site of imaging as identified by blood vessel patterns. Briefly, we anaesthetized mice with 1.5 – 3% isoflurane and used a drill (Dremel) to remove the dental cement holding the window in place. We removed the glass window was removed and injected the red marker into the site of imaging (Red Retrobeads: 250 nl, CTB: 500 nl (0.5%), mCherry: 500 nl) using a glass pipette (tip diameter: 40 – 50 µm) at 60 nl min^{-1}. Following the injection, we covered the exposed brain with silicon (Kwik-Sil, World Precision Instruments) and then coated it with dental cement. After allowing time for retrograde transport (retrobeads and CTB: 1 week) or viral transfection and expression (mCherry: 3 weeks) mice were deeply anesthetized with a mixture of Dexmedatomidine (3 mg/kg) and Ketamine (300 mg/kg) and brains were extracted following perfusion in 0.01 M phosphate buffer pH 7.4 (PBS) and 4% paraformaldehyde (PFA). They were further fixed in PFA overnight and cryopreserved in 30% sucrose solution for 2 days prior to slicing. The location of imaging was confirmed through fluorescent imaging (Figure S1). For retrobeads and CTB, the injection site was clear as a very bright injection site, for mCherry, expression levels were measured across the AC and the site of imaging was assumed to be the section with the strongest expression/brightest red. The identified sections were cross-referenced with the Allen Institute Mouse Brain Atlas using freely available software (Shamash et al., 2018).

Model

Neuron model and network.

We simulated cortical neural populations, MGB populations and a BLA neuronal population in a rate-based description of neural activity. We simulated $N = 10$ MGB populations. Each MGB population received $N = 10$ inputs ${x_i}^{MGB}$, i = 1..N. To model the fact that neighboring inputs are correlated, we generated the inputs x_i assuming that they each have a similar tuning to stimuli. These stimuli were modeled as 10 time-dependent activities $s_j(t)$ (which corresponded to a sound amplitude at a given frequency, j). The activity of input i was calculated by a sum of the stimulus channels, weighted with tuning strengths ${x^{MGB}}_i(t) = \sum_j {T^{MGB}}_{ij}\, s_j(t) + {x^{ctx}}_j(t)$. The input tuning was Gaussian: ${T^{MGB}}_{ij} = \left[e^{-\frac{(i-j)^2}{2\sigma^{MGB}}}\right]_+$ for i and j going from 1 to 10. $[.]_+$ means that negative values are set to zeros. The term x^{ctx} corresponds to the direct cortical feedback. The parameter σ^{MGB} regulated how broad the population response is to the sound. In the model, we assumed that MGB neural populations always have a small overlap in neural responses ($\sigma^{MGB} = 0.8$).

Similarly, we simulated $N = 10$ cortical populations are modelled as ${x^{ctx}}_i(t) = \sum_j {T^{ctx}}_{ij} {x^{MGB}}_j\,(t)$. The input tuning was also Gaussian: ${T^{ctx}}_{ij} = \frac{1}{1.8}\left[e^{-\frac{(i-j)^2}{2\sigma^{ctx}}} - I^{ctx}\right]_+$ for i and j going from 1 to 10. I^{ctx} = 0.9 was a broad inhibitory term.

In the simulations, we tested for two different values of σ^{ctx}; one corresponding to narrow tuning with a small overlap ($\sigma^{cxt} = 3$), and one corresponding to a broad tuning with a large overlap ($\sigma^{ctx} = 10$). (Note that $\sigma^{MGB} = 0.8$ was equivalent to $\sigma^{ctx} = 3$ since we did not model MGB inhibition here, $I^{MGB} = 0$). To avoid boundary effects, we had a circular boundary condition of the 10 inputs, meaning that input 1 and input 10 are neighbors.

Finally, we simulated one population in the BLA. It received inputs from both cortical and MGB populations, i.e. y = $w^{MGB}x^{MGB} + w^{ctx}x^{ctx}$,where w^{MGB} are the weights from MGB neurons to the BLA

neurons, and w^{ctx} are the weights from cortical neurons to the BLA. Normalized freezing response was computed as the activity after the fear conditioning paradigm (see below) normalized by the maximal activity (i.e. when the weights are all 1).

Modelling fear conditioning paradigm and interventions

During the fear conditioning training to simulate a CS- tone, we set (channel number 6) $s_6 = 1$, all the other inputs to zero, and a CS+ we set (channel number 3) $s_3 = 1$, all the other inputs to zero. In addition, we paired it with a shock (e = 1 if there is a shock, e = 0 otherwise). The synaptic weights were plastic under the following rules, $\Delta w^{ctx/MGB}{}_i = \alpha\, x^{ctx/MBG}{}_i e$, where $\alpha = 0.1$ is the learning rate. This is analogous to the standard Delta rule. The weights were bound between 0 and 1 and are initialized at 0.1. We simulated the fear conditioning for 10 time-steps [arbitrary time]. To simulate optogenetic inactivation of PV neurons in AC (Aizenberg et al., 2015), which decreases inhibition in AC, we lowered inhibition in AC by setting I^{ctx} = 0.45 (half the 'normal' level), the maximum freezing was computed with the original inhibitory term intact ($I^{ctx} = 0.9$). To simulate pharmacological inactivation of AC during memory recall (after learning), we tested the behavior of the model with AC inactivation by setting $x^{ctx}{}_i = 0$ during the BLA simulation protocol.

Data availability

All data and the code to generate the figures as well as the model will be available upon peer-reviewed publication in free access here: https://doi.org/10.5061/dryad.wpzgmsbhw.

Supplemental Information (7 figures and 6 tables)

Figure S1: Location of imaging site.

Figure S2: Neither Zdiff nor SVM performance predicts Learning Specificity post-DFC.

Figure S3: Change in Zdiff in tracked cells between consecutive sessions.

Figure S4: Model process.

Figure S5: Model process with PV inactivation of AC during DFC. The

Figure S6: Model results with AC inactivation during memory recall.

Figure S7: Auditory Brainstem Responses.

Table S1: Statistics for Figure 2C.

Table S2: Statistics for Figure 5A, paired t-tests. FDR = false discovery rate.

Table S3: Statistics for Figure 5B, paired t-tests. FDR = false discovery rate.

Table S4: Statistics for Figure 5C, paired t-tests. FDR = false discovery rate.

Table S5: Paired t-test between the mean Zdiff in session n and session n+1 after training on session n (Statistics for Figure 7D).

Table S6: Paired t-test between the mean Zdiff in session n and session n+1 after training on session n (Statistics for Figure S3).

References

Aizenberg, M., and Geffen, M.N. (2013). Bidirectional effects of aversive learning on perceptual acuity are mediated by the sensory cortex. Nat. Neurosci. *16*, 994–996.

Aizenberg, M., Mwilambwe-Tshilobo, L., Briguglio, J.J., Natan, R.G., and Geffen, M.N. (2015). Bidirectional Regulation of Innate and Learned Behaviors That Rely on Frequency Discrimination by Cortical Inhibitory Neurons. PLOS Biol. *13*, e1002308.

Aizenberg, M., Rolón-Martínez, S., Pham, T., Rao, W., Haas, J.S., and Geffen, M.N. (2019). Projection from the Amygdala to the Thalamic Reticular Nucleus Amplifies Cortical Sound Responses. Cell Rep. *28*, 605-615.e4.

Antunes, R., and Moita, M.A. (2010). Discriminative auditory fear learning requires both tuned and nontuned auditory pathways to the amygdala. J. Neurosci. *30*, 9782–9787.

Apergis-Schoute, A.M., Dębiec, J., Doyère, V., LeDoux, J.E., and Schafe, G.E. (2005). Auditory fear conditioning and long-term potentiation in the lateral amygdala require ERK/MAP kinase signaling in the auditory thalamus: A role for presynaptic plasticity in the fear system. J. Neurosci. *25*, 5730–5739.

Armony, J.L., Servan-Schreiber, D., Romanski, L.M., Cohen, J.D., and LeDoux, J.E. (1997). Stimulus generalization of fear responses: effects of auditory cortex lesions in a computational model and in rats. Cereb. Cortex *7*, 157–165.

Atencio, C., Sharpee, T., and Schreiner, C. (2012). Receptive field dimensionality increases from the auditory midbrain to cortex. J. Neurophysiol.

Bakin, J.S., and Weinberger, N.M. (1990). Classical conditioning induces CS-specific receptive field plasticity in the auditory cortex of the guinea pig. Brain Res. *536*, 271–286.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Ser. B *57*, 289–300.

Betley, J.N., Cao, Z.F.H., Ritola, K.D., and Sternson, S.M. (2013). Parallel, Redundant Circuit Organization for Homeostatic Control of Feeding Behavior. Cell *155*, 1337–1350.

Blackwell, J.M., and Geffen, M.N. (2017). Progress and challenges for understanding the function of cortical microcircuits in auditory processing. Nat. Commun. *8*, 1–9.

Boatman, J.A., and Kim, J.J. (2006). A thalamo-cortico-amygdala pathway mediates auditory fear conditioning in the intact brain. Eur. J. Neurosci. *24*, 894–900.

Briguglio, J.J., Aizenberg, M., Balasubramanian, V., and Geffen, M.N. (2018). Cortical Neural Activity Predicts Sensory Acuity Under Optogenetic Manipulation. J. Neurosci. *38*, 2094–2105.

Ceballo, S., Piwkowska, Z., Bourg, J., Daret, A., and Bathellier, B. (2019). Targeted Cortical Manipulation of Auditory Perception. Neuron *104*, 1168-1179.e5.

Chapuis, J., and Wilson, D. a (2011). Bidirectional plasticity of cortical pattern recognition and behavioral sensory acuity.

Chen, L., Wang, X., Ge, S., and Xiong, Q. (2019a). Medial geniculate body and primary auditory cortex differentially contribute to striatal sound representations. Nat. Commun. *10*, 1–10.

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.

Chen, X., Sun, Y.-C., Zhan, H., Kebschull, J.M., Fischer, S., Matho, K., Huang, Z.J., Gillis, J., and

Zador, A.M. (2019b). High-Throughput Mapping of Long-Range Neuronal Projection Using In Situ Sequencing. Cell *179*, 772-786.e19.

Dalmay, T., Abs, E., Poorthuis, R.B., Hartung, J., Pu, D.-L., Onasch, S., Lozano, Y.R., Signoret-Genest, J., Tovote, P., Gjorgjieva, J., et al. (2019). A Critical Role for Neocortical Processing of Threat Memory. Neuron 1–15.

DiCarlo, J.J., Zoccolan, D., and Rust, N.C. (2012). How does the brain solve visual object recognition? Neuron *73*, 415–434.

Dunsmoor, J.E., and Paz, R. (2015). Fear Generalization and Anxiety: Behavioral and Neural Mechanisms. Biol. Psychiatry *78*, 336–343.

Edeline, J.-M. (1999). Learning-induced physiological plasticity in the thalamo-cortical sensory systems: a critical evaluation of receptive field plasticity, map changes and their potential mechanisms. Prog. Neurobiol. *57*, 165–224.

Edeline, J.M., and Weinberger, N.M. (1993). Receptive field plasticity in the auditory cortex during frequency discrimination training: selective retuning independent of task difficulty. Behav. Neurosci. *107*, 82–103.

Georgopoulos, A.P., Schwartz, A.B., and Kettner, R.E. (1986). Neuronal Population Coding of Movement Direction.

Gillet, S.N., Kato, H.K., Justen, M.A., Lai, M., and Isaacson, J.S. (2018). Fear learning regulates cortical sensory representations by suppressing habituation. Front. Neural Circuits *11*, 112.

Ginat-Frolich, R., Klein, Z., Katz, O., and Shechner, T. (2017). A novel perceptual discrimination training task: Reducing fear overgeneralization in the context of fear learning. Behav. Res. Ther. *93*, 29–37.

Groppe, D. (2020). fdr_bh (https://www.mathworks.com/matlabcentral/fileexchange/27418-fdr_bh). MATLAB Cent. File Exch.

Harris, K.D., and Shepherd, G.M.G. (2015). The neocortical circuit: themes and variations. Nat. Neurosci. *18*, 170–181.

He, J. (2003). Corticofugal modulation of the auditory thalamus. Exp. Brain Res. *153*, 579–590.

Herry, C., and Johansen, J.P. (2014). Encoding of fear learning and memory in distributed neuronal circuits. Nat. Neurosci. *17*, 1644–1654.

Johansen, J.P., Cain, C.K., Ostroff, L.E., and LeDoux, J.E. (2011). Molecular Mechanisms of Fear Learning and Memory. Cell *147*, 509–524.

Jovanovic, T., and Ressler, K.J. (2010). How the neurocircuitry and genetics of fear inhibition may inform our understanding of PTSD. Am. J. Psychiatry *167*, 648–662.

Kato, H.K., Gillet, S.N., and Isaacson, J.S. (2015). Flexible Sensory Representations in Auditory Cortex Driven by Behavioral Relevance. Neuron *88*, 1027–1039.

Kato, H.K., Asinof, S.K., and Isaacson, J.S. (2017). Network-Level Control of Frequency Tuning in Auditory Cortex. Neuron *95*, 412-423.e4.

Krusemark, E.A., and Li, W. (2012). Enhanced olfactory sensory perception of threat in anxiety: An event-related fMRI study. Chemosens. Percept. *5*, 37–45.

Lange, I., Goossens, L., Michielse, S., Bakker, J., Lissek, S., Papalini, S., Verhagen, S., Leibold, N., Marcelis, M., Wichers, M., et al. (2017). Behavioral pattern separation and its link to the neural

mechanisms of fear generalization. Soc. Cogn. Affect. Neurosci. *12*, 1720–1729.

Ledoux, J.E. (2000). EMOTION CIRCUITS IN THE BRAIN.

Lee, C.C. (2015). Exploring functions for the non-lemniscal auditory thalamus. Front. Neural Circuits *9*, 1–8.

Letzkus, J., Wolff, S., Meyer, E.M., Tovote, P., Courtin, J., Herry, C., and Lüthi, A. (2011). A disinhibitory microcircuit for associative fear learning in the auditory cortex. Nature *480*, 331–336.

Li, W., Howard, J.D., Parrish, T.B., and Gottfried, J.A. (2008). Aversive learning enhances perceptual and cortical discrimination of indiscriminable odor cues. Science (80-.). *319*, 1842–1845.

Linden, J.F., and Schreiner, C.E. (2003). Columnar Transformations in Auditory Cortex? A Comparison to Visual and Somatosensory Cortices. Cereb. Cortex *13*, 83–89.

Mahan, A.L., and Ressler, K.J. (2012). Fear conditioning, synaptic plasticity and the amygdala: Implications for posttraumatic stress disorder. Trends Neurosci. *35*, 24–35.

O'Sullivan, C., Weible, A.P., and Wehr, M. (2019). Auditory Cortex Contributes to Discrimination of Pure Tones. Eneuro *6*, ENEURO.0340-19.2019.

Ohl, F.W., and Scheich, H. (1996). Differential frequency conditioning enhances spectral contrast sensitivity of units in auditory cortex (field Al) of the alert Mongolian gerbil. Eur. J. Neurosci. *8*, 1001–1017.

Ohl, F.W., and Scheich, H. (2004). Fallacies in behavioural interpretation of auditory cortex plasticity. Nat. Rev. Neurosci. *5*, 1–1.

Pachitariu, M., Stringer, C., Schröder, S., Dipoppa, M., Rossi, L.F., Carandini, M., Harris, K.D., Callaway, E.M., Keller, G., Rózsa, B., et al. (2017). Suite2p: beyond 10,000 neurons with standard two-photon microscopy (Cold Spring Harbor Labs Journals).

Quirk, G.J., Armony, J.L., and LeDoux, J.E. (1997). Fear conditioning enhances different temporal components of tone-evoked spike trains in auditory cortex and lateral amygdala. Neuron *19*, 613–624.

Resnik, J., Sobel, N., and Paz, R. (2011). Auditory aversive learning increases discrimination thresholds. Nat. Neurosci. *14*, 791–796.

Roesmann, K., Wiens, N., Winker, C., Rehbein, M.A., Wessing, I., and Junghoefer, M. (2020). Fear generalization of implicit conditioned facial features – Behavioral and magnetoencephalographic correlates. Neuroimage *205*, 116302.

Romanski, L.M., and LeDoux, J.E. (1992). Bilateral destruction of neocortical and perirhinal projection targets of the acoustic thalamus does not disrupt auditory fear conditioning. Neurosci. Lett. *142*, 228–232.

Rust, N.C., and DiCarlo, J.J. (2012). Balanced Increases in Selectivity and Tolerance Produce Constant Sparseness along the Ventral Visual Stream. J. Neurosci. *32*, 10170–10182.

Shamash, P., Carandini, M., Harris, K., and Steinmetz, N. (2018). A tool for analyzing electrode tracks from slice histology. BioRxiv 447995.

Stringer, C., Pachitariu, M., Steinmetz, N., Reddy, C.B., Carandini, M., and Harris, K.D. (2019). Spontaneous behaviors drive multidimensional, brainwide activity. Science (80-.). *364*.

Suga, N. (2008). Role of corticofugal feedback in hearing. J. Comp. Physiol. A *194*, 169–183.

Taylor, J.A., Hasegawa, M., Benoit, C.M., Freire, J.A., Theodore, M., Ganea, D.A., Lu, T., and Gründemann, J. (2020). Single cell plasticity and population coding stability in auditory thalamus upon

associative learning. BioRxiv.

Tuominen, L., Boeke, E., DeCross, S., Wolthusen, R.P., Nasr, S., Milad, M., Vangel, M., Tootell, R., and Holt, D. (2019). The relationship of perceptual discrimination to neural mechanisms of fear generalization. Neuroimage *188*, 445–455.

Weinberger, N.M. (2004). Specific long-term memory traces in primary auditory cortex. Nat. Rev. Neurosci. *5*, 279–290.

Weinberger, N.M. (2011). The medial geniculate, not the amygdala, as the root of auditory fear conditioning. Hear. Res. *274*, 61–74.

Weinberger, N.M., and Diamond, D.M. (1987). Physiological plasticity in auditory cortex: Rapid induction by learning. Prog. Neurobiol. *29*, 1–55.

Wigestrand, M.B., Schiff, H.C., Fyhn, M., LeDoux, J.E., and Sears, R.M. (2017). Primary auditory cortex regulates threat memory specificity. Learn. Mem. *24*, 55–58.

Williamson, R.S., and Polley, D.B. (2019). Parallel pathways for sound processing and functional connectivity among layer 5 and 6 auditory corticofugal neurons. Elife *8*.

Wood, K.C., Blackwell, J.M., and Geffen, M.N. (2017). Cortical inhibitory interneurons control sensory processing. Curr. Opin. Neurobiol. *46*, 200–207.

Zhang, G.-W., Sun, W., Zingg, B., Shen, L., He, J., Xiong, Y., Tao, H.W., and Zhang, L.I. (2018). A Non-canonical Reticular-Limbic Central Auditory Pathway via Medial Septum Contributes to Fear Conditioning. Neuron *97*, 406-417.e4.