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 and corresponding Author:

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

Abstract

Learning to avoid dangerous signals while preserving normal responses to safe stimuli is essential for everyday behavior and survival. Fear learning has a high level of inter-subject variability. Following identical experiences, subjects exhibit fear specificities ranging from high (specializing fear to only the dangerous stimulus) to low (generalizing fear to safe stimuli). Pathological fear generalization underlies emotional disorders, such as post-traumatic stress disorder. The neuronal basis of fear specificity remains unknown. Here, we identified the neuronal code that underlies inter-subject variability in fear specificity using longitudinal imaging of neuronal activity before and after differential fear conditioning in the auditory cortex of mice. Neuronal activity prior to, but not after learning predicted the level of specificity following fear conditioning across subjects. Stimulus representation in auditory cortex was reorganized following conditioning. However, the reorganized neuronal activity did not relate to the specificity of learning. These results present a novel neuronal code that determines 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 [1,2]. However, the same fear conditioning paradigm elicits different levels of learning specificity across subjects [3–6]. 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) [7,8] and anxiety [9]. 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) is involved in fear learning. *During* differential fear conditioning (DFC), inactivation of AC chemically [10], or with optogenetics [11], as well as partial suppression of inhibition in AC [12] led to decreased learning specificity using either pure tones or complex stimuli, such as FM sweeps or vocalizations [3,11–14]. These observations suggest that AC may determine the level of learning specificity. Therefore, we tested whether neuronal codes in AC *prior* to conditioning can predict specificity of fear learning.

The role of AC *following* fear conditioning is more controversial. Changes in stimulus representation in AC following association learning have been proposed to represent multiple different features of the fear response [1,14–19]. However, inactivation of the auditory cortex did not affect fear memory retrieval of pure tones [3,11], suggesting that AC is not involved in fear memory retrieval. If AC were involved in fear memory retrieval, we would expect the changes in sound representation to reflect the level of learning specificity across subjects. Therefore, our second goal was to test the role of changes in auditory cortex in shaping fear learning specificity across subjects.

To address these goals, we imaged the activity of neuronal ensembles in layers 2 and 3 of AC over weeks, before and after differential fear conditioning with pure tones. 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 changes in stimulus representation 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 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

Learning specificity varies amongst conditioned mice.

To establish the relationship between sound-evoked activity in the AC and differential fear conditioning, we recorded simultaneous neuronal activity from hundreds of neurons in AC. We tracked the same neurons before and after DFC, using two-photon imaging of a fluorescent calcium probe (GCaMP6 [20], Fig. S1-2). Longitudinal imaging of neuronal activity in large ensembles of neurons in layers 2 and 3 of AC before and after conditioning allowed us to compare the representation of the CS stimuli before and after learning (Fig. 1A).

Mice exhibited a range of learning specificities. We conditioned mice by exposure to 10 repeats of an alternating sequence 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, but the foot-shock occurred during periods of silence between the stimuli (Fig. 1B). We measured fear-memory retrieval by presenting the same auditory stimuli to the mice in a different context and measuring the % of time the mice froze during stimulus presentation and at baseline (Fig. 1C). Following conditioning, memory retrieval was tested after each imaging session and the levels of freezing were consistent across all 4 retrieval sessions (Fig 1D-E, two-way repeated-measures (rm)-ANOVA, no effect of retrieval session on freezing, $F_{(3, 126)} = 0.91$, $p = .440$. Nor any interaction between session and stimulus type (session*baseline/CS-/CS+, $F_{(6,126)} = 0.70$, $p = .651$). Similarly, freezing in *pseudo*-conditioned mice was consistent over the 4 retrieval sessions (two-way rm-ANOVA, no main of session on freezing, $F_{(3,90)} = 0.90$, $p = .442$; no interaction between session and stimulus type ($F_{(6,90)} = 0.66$, $p = .683$). Since there was no change in freezing over time, we do not specifically consider results with respect to the second DFC session (Fig. 1A, day 12). Henceforth we refer to DFC as the first DFC session. Conditioned mice that did not freeze to CS+ or CS- differently from baseline across all 4 retrieval sessions were excluded from subsequent analysis (6/21 mice excluded, Fig. S3A, two-way ANOVA, $p > .05$, see methods).

Learning specificity was defined as the difference between freezing to CS+ and CS- during memory retrieval sessions (see Methods, Equation 1) [3]. We used two pure-tone CS stimuli which have been shown to engage AC in human DFC [21]. The pure tones were close together in frequency space (0.40 octaves apart) in order to drive a range of learning specificities in conditioned mice that are not achievable at greater frequency distances [3]. Indeed, we observed that conditioned mice displayed a larger range of learning specificities (range: -16.9 to 55.6%) compared with *pseudo*-conditioned mice (-5.6 to 9.1%). This was reflected in a significantly larger standard deviation of learning specificity in conditioned mice ($\sigma = 20.2\%$) than in *pseudo*-conditioned mice (Fig. 1F, $\sigma = 4.5\%$, *F*-test, $F_{(14, 10)} = 20.48$, $p < .001$) in the first retrieval session after DFC. We also observed a significantly higher learning specificity (mean: 14.5%) in conditioned mice than *pseudo*-conditioned mice (mean: 1.1%, *t*-test, $t_{(24)} = 2.15$, $p = .042$) in the retrieval session after DFC. Learning specificity persisted for the subsequent memory retrieval sessions over the course of the experiment in conditioned and *pseudo*-conditioned mice (Fig. S3B-D, two-way rm-ANOVA, $p > 0.05$, Table S1). Thus, we found that conditioned mice exhibited a range of learning specificities, with some generalizing their fear across the CS stimuli and others specializing their fear responses to CS+. On average, the learning specificity of mice was stable over the course of the experiment.

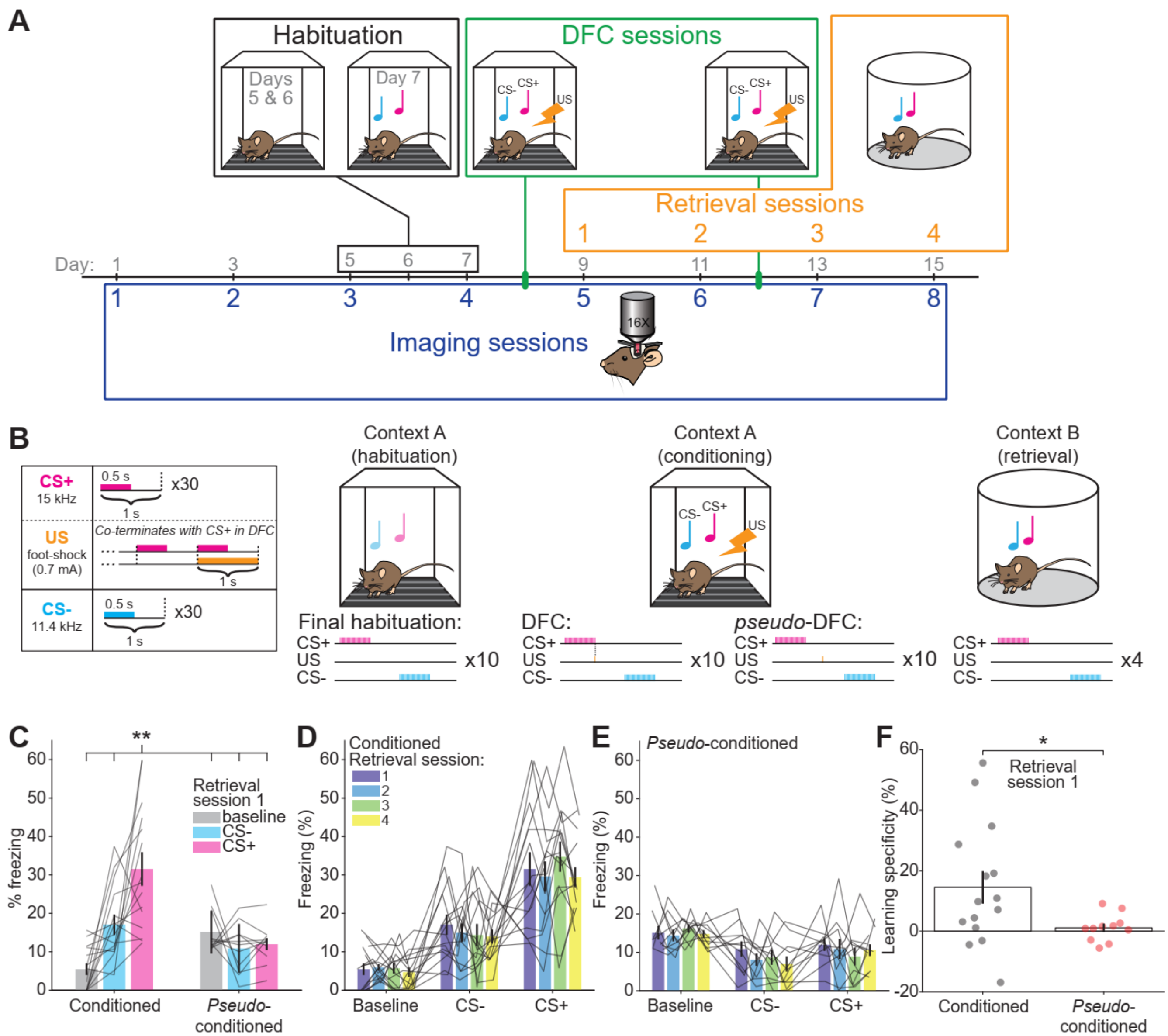

Figure 1: Experimental timeline and differential fear conditioning (DFC) paradigm. (A) Experimental timeline: Mice were imaged for 4 sessions (48 hours apart) before DFC to establish baseline responses to tone pip stimuli under the two-photon. Prior to DFC, mice were habituated to the fear conditioning chamber. Mice were subjected to DFC (21 mice) or *pseudo*-conditioning (11 mice) on Days 8 and 12. After DFC-1 (day 8), fear retrieval testing was performed after each imaging session. **(B)** Mice were habituated to the conditioning chamber (context A) for 3 days prior to conditioning and on the final day, the stimuli were presented without foot-shock. During conditioning, a foot-shock (1 s, 0.7 mA) was paired with the CS+ (15 kHz, 30s pulsed at 1 Hz). The CS- (11.4 kHz, 30 s pulsed at 1 Hz) was presented alternately with the CS+ (30-180 s apart, 10 repeats) and not paired with a foot-shock. During *pseudo*-conditioning, 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). Motion of the mouse was recorded and the percentage freezing during each stimulus was measured offline. **(C)** Freezing at baseline (gray), for CS+ (pink) and CS- (blue) in retrieval session 1 (day 9) showing the

percentage of time frozen during tone presentation for CS+, CS- and baseline for each mouse. Gray lines indicate freezing for each included mouse. (** Two-way ANOVA (Table S2), *Tukey-Kramer post-hoc* test, $p < .01$). **(D)** Freezing to baseline, CS-, and CS+ for each conditioned mouse over the 4 retrieval sessions. Gray lines show each mouse. **(E)** Same as **D** for *pseudo*-conditioned mice. **(F)** Learning specificity of conditioned and *pseudo*-conditioned mice for retrieval session 1. Circles show individual mice. (**t*-test, $p < .05$). Error bars in **C-F** indicate standard error of the mean (SEM).

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

We used two-photon imaging to record calcium activity from neurons in auditory cortex in head-fixed mice (Fig. 2A). We presented 100-ms tone pips (frequency range: 5-32 kHz, including CS+ and CS- frequencies) to obtain frequency response functions from each neuron. We hypothesized that the activity in auditory cortex would predict learning specificity across individual mice. Thus, we tested whether neuronal discrimination of CS+ and CS- in AC pre-DFC predicted learning specificity following DFC. To assess how well single neurons could discriminate between the two conditioned tones, we computed the Z-score difference (Z_{diff}, see Methods, Equation 2) of responses to CS+ and CS- for responsive neurons. In an example neuron (Fig. 2B), the distributions of single-trial response magnitudes to CS+ and CS- demonstrate a separation resulting in a significant Z_{diff} score of 2.01. The Z_{diff} score of responsive neurons was considered significant if the actual score was greater than the 95th percentile of the bootstrapped Z_{diff} scores (see Methods). Figure 2C shows the distribution of Z_{diff} scores for all responsive units from conditioned mice 24 hours pre-DFC. We found that the mean significant Z_{diff} scores across the 4 pre-DFC imaging sessions predicted the learning specificity 24 hours post-DFC (Fig. 2D, $r(13) = .74$, 95% CI [.51, .88], $p = .002$). In fact, this prediction was already evident using Z_{diff} from session 4 alone (Fig. 2E, $r(12) = .62$, 95% confidence intervals (CI) [.00, .83], $p = .042$, one mouse did not have any neurons with significant Z_{diff} in this session, correlation not different from using mean Z_{diff} across all pre-DFC sessions, bootstrap comparison, see methods: r difference = 0.11 , 95% CI [-0.10, 0.65], $p = 0.316$, $N = 14$). In summary, this suggests that the individual neuronal discriminability in AC pre-DFC predicts the learning specificity 24 hours post-DFC.

It is possible that the Z_{diff} score results from some underlying distributions of response magnitudes; for example, the magnitude of response to CS+ could be the driver of the prediction phenomenon. Thus, we explored whether magnitude of CS+ or CS- responses related to learning specificity. We compared the mean response magnitudes to each CS over the 4 pre-DFC imaging sessions with learning specificity 24 hours post-DFC and found that the were not correlated (Pearson's correlation, $p < 0.05$, Fig. S4). This suggests that it is not merely the magnitude of responses to CS+ or CS- but truly discriminability of the responses that is underlying the prediction of learning specificity.

We next tested the temporal window for the prediction of learning specificity. If changes in sound-evoked responses in AC reflect memory formation or the strength of learning as previously suggested [15,17], we would expect a stronger relationship between the neuronal discrimination and learning specificity after DFC than before. To test this, we compared the correlations between mean Z_{diff} across equal numbers of imaging sessions (3 imaging sessions preceding retrieval sessions 1, and 4) and learning specificity before and after DFC. We found that the mean Z_{diff} score pre-DFC predicted learning specificity from retrieval session 1 ($r(13) = .73$, 95% CI [.43, .88], $p = .003$), whereas the mean Z_{diff} score post-DFC did not predict learning specificity in retrieval session 4 ($r(13) = .108$, 95% CI [-.37 .61], $p = .727$). Furthermore, the correlation pre-DFC was significantly stronger than that post-DFC (bootstrap comparison (see Methods): r difference = 0.626, 95% CI [0.04, 1.14], $p = .036$).

In summary, individual neuronal discriminability in AC pre-DFC predicted learning specificity 24 hours after DFC. Post-DFC, neuronal activity no longer predicted learning specificity. Therefore, the role of auditory cortex in DFC is likely restricted temporally. To further investigate the relationship between neuronal and behavioral discriminability, we examined whether neuronal population discriminability could predict learning specificity.

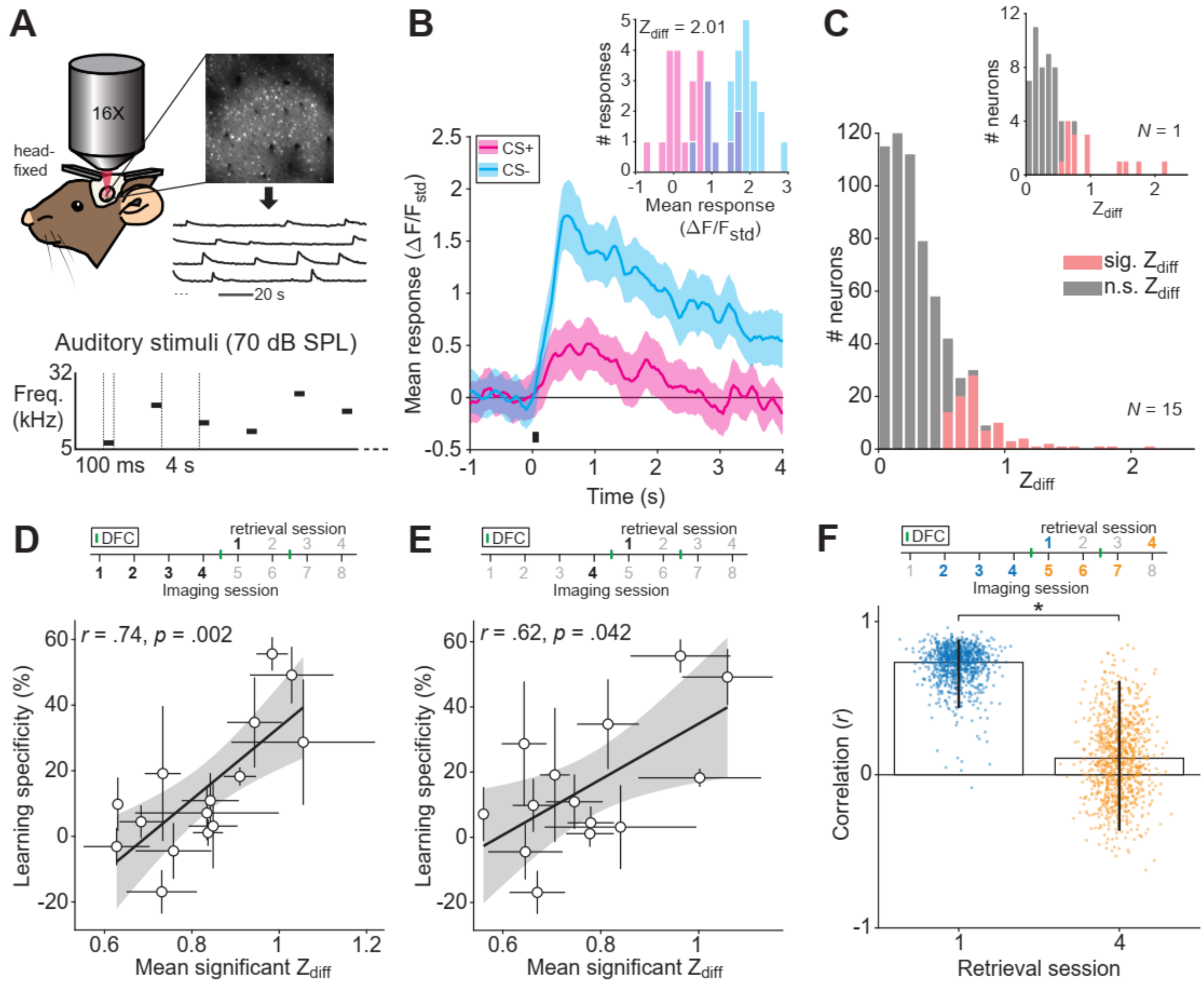

Figure 2: Mean neuronal discriminability pre-DFC predicts learning specificity. (A) Imaging setup: Mice were head-fixed under the two-photon microscope, fluorescence of calcium indicator (GCaMP6s/m) was measured at ~30 Hz, regions of interest and mean fluorescence over time were extracted using open software [22]. Schematic showing auditory stimuli, comprised of pure-tone pips (100 ms, 70 dB SPL, 5-32 kHz) presented at 0.24 Hz. **(B)** 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 (mean $\Delta F/F_{std}$ across 2-s window following stimulus onset) to CS+ and CS- from the same neuron. **(C)** Distribution of Z_{diff} scores of responsive units from conditioned mice 24 hours pre-DFC. Significant scores are indicated in red, *N* = 94/617 neurons. Inset, single mouse example, *N* = 15/62 neurons. **(D)** Mean Z_{diff} score of significant-Z_{diff} neurons for each mouse 24 hours pre-DFC (imaging session 4) does not correlate with learning specificity 24 hours post-DFC (retrieval session 1). Error bars = SEM; black line = linear regression between the mean Z_{diff} and learning specificity, gray shading = 95% CI of the fit. (**E**) Mean Z_{diff} score of significant-Z_{diff} neurons pre-DFC (imaging sessions 1-4) correlated with learning specificity 24 hours post-DFC. **(F)** Pearson's correlation (*r* ± 95% CI) between Z_{diff} score averaged across 3 imaging sessions preceding retrieval sessions 1 (blue),

and 4 (orange). Dots represent individual bootstrapped (n = 1000) correlation values. * $p < 0.05$, bootstrap comparison, see methods.

Population neuronal activity in AC pre-DFC predicts specificity of fear learning.

For many brain regions and tasks, activity of multiple neurons can provide more information in combination than averaged activity of individual neurons [23–25]. Using machine learning, we investigated whether populations of neurons predicted learning specificity better than the average Z_{diff} scores. We trained a Support Vector Machine (SVM) to discriminate between presentation of CS+ and CS- using population responses to the two stimuli. Mean SVM performance prior to DFC correlated with learning specificity 24 hours post-DFC (Fig. 3A, $r(13) = .85$, 95% CI [.66, .94], $p < .001$). However, unlike with the mean Z_{diff}, SVM performance 24 hours pre-DFC did not predict learning specificity 24 hours post-DFC (Fig. 3B, $r(13) = .43$, 95% CI [.05, .76], $p = .109$); significantly weaker than that of all pre-DFC imaging sessions (bootstrap comparison (see Methods), r difference = -0.42, 95% CI [-0.71, -0.13], $p = .004$), potentially due to the SVM including neurons with non-specific responses that likely differed across imaging sessions. Nonetheless, Z_{diff} scores and the SVM performance of the same neurons were strongly correlated (Fig. 3C, $r(13) = .88$, 95% CI [.77, .95], $p < 0.001$), suggesting that the two different discriminability methods used similar underlying features to discriminate the stimuli. This was also reflected in the fact that the correlations between the two discriminability measures across pre-DFC imaging sessions and learning specificity were not statistically different (bootstrap comparison, r difference = -0.11, 95% CI [-0. 25, 0.01], $p = .078$). Thus, population responses pre-DFC predicted subsequent learning specificity likely through similar mechanisms to the mean Z_{diff}.

We next tested whether predictability of learning specificity persisted after DFC by comparing the mean SVM performance across the 3 preceding imaging sessions with retrieval sessions 1 and 4 (Fig. 3D). The mean SVM performance pre-DFC (predicted learning specificity in retrieval session 1 ($r(13) = .78$, 95% CI [.38, .94], $p < .001$). However, the mean SVM performance post-DFC did not predict learning specificity in retrieval session 4 ($r(13) = .46$, 95% CI [-.11, .80], $p = .092$). However, the two correlations were not significantly different (bootstrap comparison, r difference = 0.331, 95% CI [-0.12, 0.90], $p = .150$). Combined with similar results from the mean Z_{diff} scores (Fig. 2F), there is some evidence to support that neuronal discriminability before conditioning predicts learning specificity but no longer predicts learning specificity after conditioning. This suggests that neuronal discrimination in the output layers of AC is involved in the initial discrimination during, but not after, differential fear conditioning.

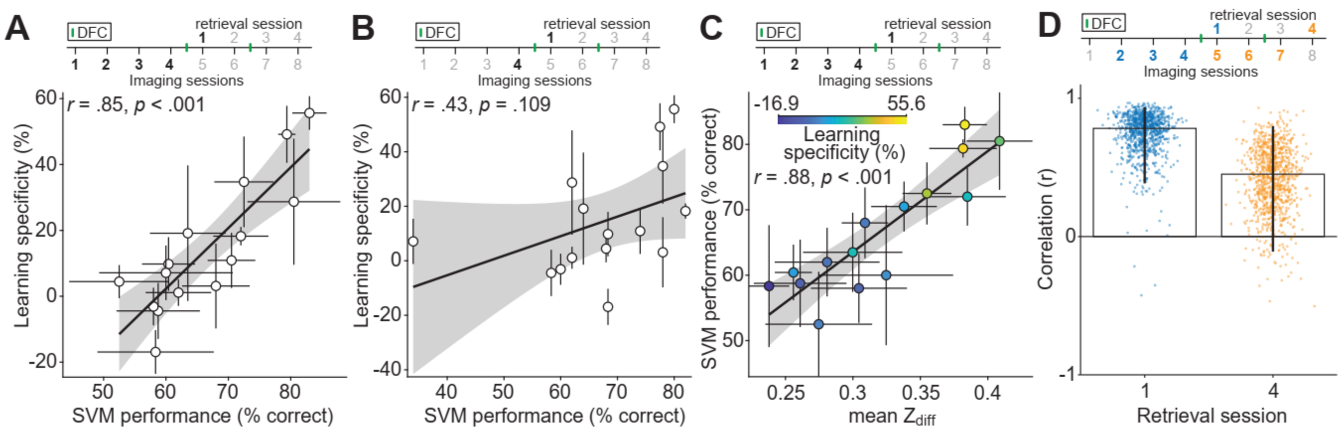

Figure 3: Neuronal population discrimination between CS+ and CS- pre-DFC predicts learning specificity. (A) SVM performance 24 hours pre-DFC (imaging session 4) does not predict learning specificity 24 hours post-DFC (retrieval session 1, N = 15). **(B)** SVM performance (Mean ± SEM) across pre-DFC sessions (imaging sessions 1-4) predicts learning specificity 24 hours post-DFC (retrieval session 1). **(C)** Mean (± SEM) SVM performance pre-DFC correlates with the mean (± SEM) Z_{diff} score pre-DFC (imaging sessions 1-4). Fill color indicates learning specificity from retrieval session 1. **(D)** Correlation (r ± 95% CI) between SVM performance averaged across 3 imaging sessions preceding retrieval sessions 1 (blue), and 4 (orange). Dots represent individual bootstrapped correlation values (n = 1000). Black lines in **A**, **B**, & **C** show the linear fit between the two variables, shading = 95% CI. Circles represent individual mice with error bars = SEM. Statistics in **A-D**: Pearson's correlation.

After DFC, discriminability between CS+ and CS- is preserved in conditioned mice.

It has been suggested that 'fear memories' are encoded in the auditory cortex following differential fear conditioning [15,17], such that neuronal discriminability may improve following conditioning. We found that neuronal activity following DFC no longer predicted learning specificity (Fig. 2F, 3D), suggesting AC does not support the fear response after DFC. We tested whether the neuronal discriminability of CS+ and CS- changed after DFC by comparing the mean Z_{diff} across pre- and post-DFC sessions (Fig. 4A, rm-ANOVA, Table S3). We found no change in Z_{diff} from pre- to post-DFC in conditioned mice (p = .581), whereas there was a significant decrease in *pseudo*-conditioned mice (-0.11 ± 0.05 units, *Tukey-Kramer post-hoc* comparison, p = .026). Results were similar at a neuronal population level; mean SVM performance in conditioned mice did not change across pre- and post-DFC sessions (Fig. 4B, rm-ANOVA, Table S4, *Tukey-Kramer post-hoc* comparison, p = .802), whereas there was a significant decrease in *pseudo*-conditioned mice (-4.7 ± 1.8%, *Tukey-Kramer post-hoc* comparison, p = .014).

Combined, we found that following DFC or *pseudo*-conditioning, neuronal discrimination between the CS+ and CS- was maintained in conditioned mice, while it decreased in *pseudo*-conditioned mice. These results suggest that AC does not store fear memory by increasing discriminability. Rather, plasticity in AC appears to counteract previously reported habituation in neuronal responses to repeated stimuli [18,26].

To further investigate whether the changes in stimulus representation were related to DFC, we tested the neuronal discrimination 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 reorganization of stimulus representation between days contributes to stimulus discrimination, we would expect a change in performance following DFC. We did not observe any consistent changes in SVM performance following DFC in conditioned or *pseudo*-conditioned mice (Fig. S5A, two-way rm-ANOVA with dependent variable; SVM performance and independent variables; session comparison (1 & 2, 2 & 3… to 7 & 8) and conditioning type (DFC/*pseudo*-conditioning, no effect of) $p > 0.05$ for session and interaction, Table S5). Similarly, we did not observe any significant changes in mean Z_{diff} in neurons tracked between any of the consecutive sessions of the experiment (Fig. S5B, 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 any correlation between the neuronal discriminability (mean Z_{diff} score and SVM performance) and the learning specificity *post*-DFC. The mean Z_{diff} score (imaging sessions 5-8) did not correlate with the mean learning specificity of conditioned mice across the 4 post-DFC sessions (Fig. 4C, Pearson's correlation, $r(13) = .05$, CI [-.45, .56], $p = .838$), nor was there a correlation between the mean SVM performance post-DFC and the mean learning specificity post-DFC (Fig. 4D, $r(13) = .49$, CI [-.19, .86], $p = .066$). This suggests that neuronal discriminability post-DFC does not reflect learning specificity.

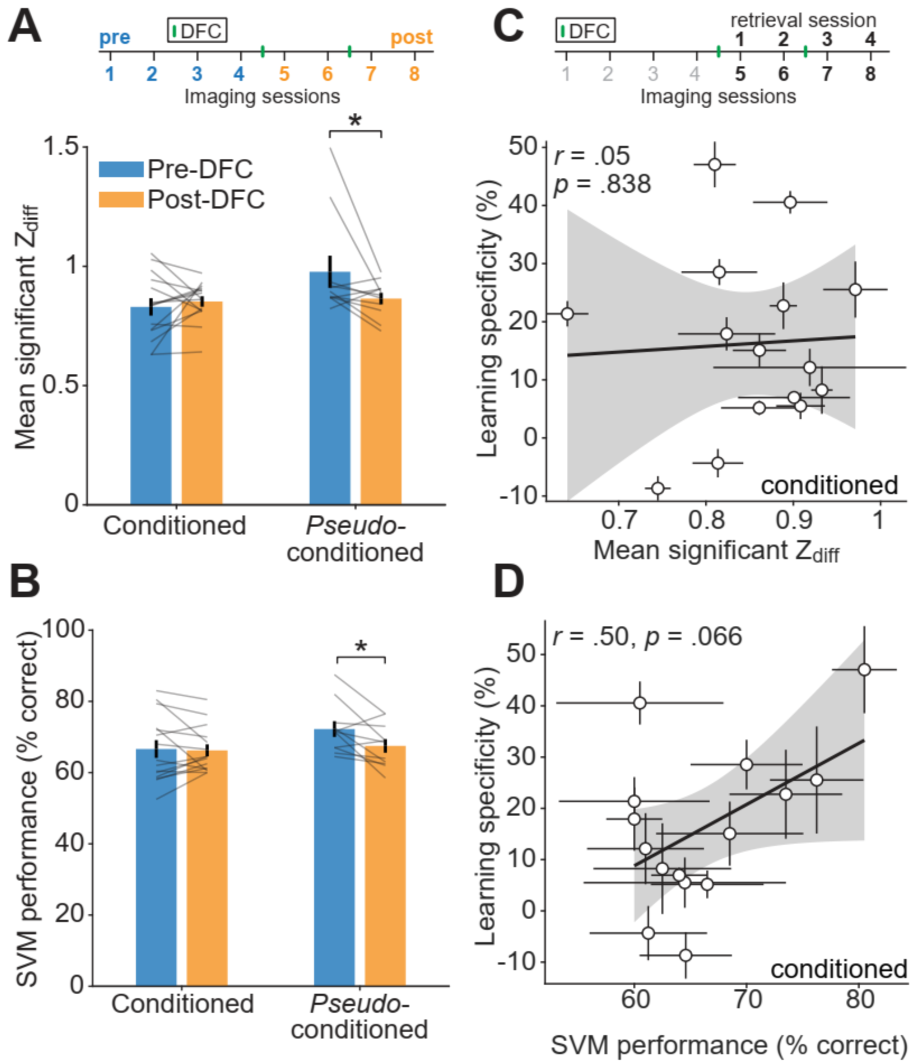

Figure 4: Changes in stimulus information representation post-DFC. (A) Comparison of mean ± SEM significant-Z_{diff} between the pre- (sessions 1-4, blue) and post-DFC sessions (5-8, orange) in conditioned and *pseudo*-conditioned mice. Stats: *Tukey-Kramer post-hoc:* *$p < .05$, Table S3. **(B)** Same as **A** but for comparison of mean ± SEM SVM performance between the pre- and post-DFC. Stats: *Tukey-Kramer post-hoc:* *$p < .05$, Table S4. **(C)** Relationship between mean significant-Z_{diff} across the post-DFC sessions (sessions 5-8) and mean learning specificity across the same sessions. Statistics: Pearson's correlation. Black line shows linear fit (shading = 95% CI of the fit). **(D)** Same as **C** but for mean SVM performance across the post-DFC sessions (sessions 5-8) and mean learning specificity across the same sessions. Statistics: Pearson's correlation.

After DFC, normalized responses at CS+ increased in conditioned mice.

It has previously been shown that after differential conditioning with pure tones, select neurons in AC amplified the difference between CS+ and CS- [17,27]. However, since we observed no change in neuronal discrimination in conditioned mice, we hypothesized that there would be no change in response to CS+ and CS-. To test whether responses were altered by conditioning, we compared frequency response functions from the pre- and post-DFC imaging sessions of responsive neurons that were tracked from pre- to post-DFC (Fig. S6A). On an individual neuron basis, we observed heterogeneous changes in the frequency tuning (Fig. 5A, Table S7-9). However, we found that, on average, in conditioned mice, the *normalized* response to CS+ and frequencies between the CS+ and CS- increased, whereas the response at CS- did not change (Fig. 5B, two-way rm-ANOVA, *Tukey-Kramer post-hoc* testing, $p < .05$, Table S10). In contrast, in *pseudo*-conditioned mice, the mean normalized responses at most frequencies, including both CS frequencies, did not change (Fig. 5C, Table S11). When comparing responses at CS- and CS+ in conditioned mice and the CS stimuli combined (CSc) in *pseudo*-conditioned mice, we found that there was a significant increase at the CS+ and no change at CS- or CSc (Table S12, *Tukey-Kramer post-hoc* comparison, $p = .002$). Although we observed an increase in normalized response to CS+, we did not observe any significant changes in non-normalized response to conditioned frequencies in conditioned mice (Fig. S6B, D, & E, Table S13-15) and we observed decreased responses to most frequencies in *pseudo*-conditioned mice (Fig. S6C, F, & G, Table S16-18). When comparing non-normalized response changes to CS+, CS- and CSc, we found a significant decrease at CSc but not at CS+ or CS- (Table S19, *Tukey-Kramer post-hoc* comparison, $p < .001$). It is possible that the normalization of the frequency response functions has amplified a small change that is not strong enough to present in the absolute responses.

Despite the lack of significant change in the absolute responses, it is possible that the increase in normalized responses at CS+ and the lack of change in response at CS- in conditioned mice could lead to improved discriminability between CS+ and CS- by increasing the difference between the responses to each stimulus. This would be consistent with the hypothesis that, following fear conditioning, reorganization of neuronal activity serves to amplify the *relative* difference in responses to CS+ and CS- thereby supporting discriminability [17,19]. However, when we compared the magnitude of changes in normalized response to CS+, CS-, and the difference between the two with learning specificity, we did not find any correlation (Fig. S7), suggesting that the changes observed are in fact not related to storage of the fear memory [28]. We further investigated by checking for a relationship between the *change* in neuronal discrimination (Z_{diff} & SVM performance) and learning specificity (Fig. S8) finding negative correlations between the two factors. This possibly supports the idea that cortical activity is reorganized following DFC in a way that does not support this form of fear memory.

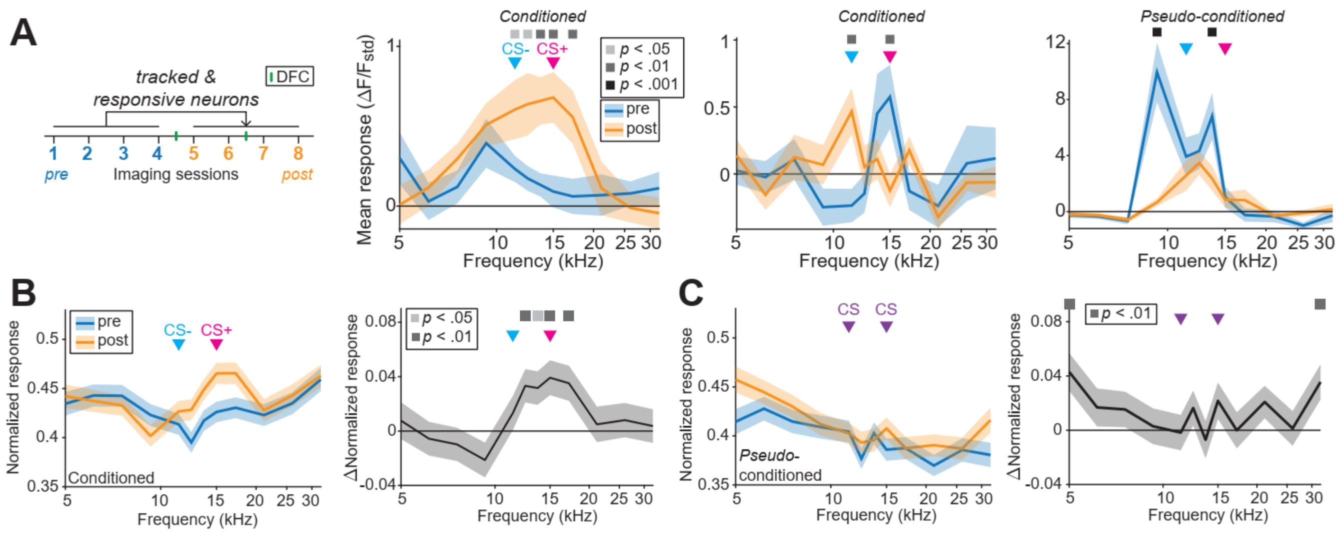

Figure 5: Changes in frequency representation post-DFC. (A) We tracked the responses of the same responsive neurons pre- and post-DFC. The right panels show three example frequency response functions from tracked neurons from conditioned and *pseudo*-conditioned mice pre-DFC (blue) and post-DFC (orange). Significant differences in the response functions are indicated by the squares above (paired *t*-test, FDR-corrected for multiple comparisons – see methods). The two arrows show the frequencies of the CS- (11.4 kHz) and CS+ (15 kHz) and CSc. Two-way rm-ANOVA, *Tukey-Kramer post-hoc* analysis (Table S7-9). **(B)** (left panel) Mean normalized frequency response functions of responsive neurons that were tracked pre- to post-DFC across all conditioned mice (N = 15 mice, N = 826 neurons). (right panel) Mean change in normalized frequency response functions of the same neurons, squares indicate significant changes (Two-way rm-ANOVA, *Tukey-Kramer post-hoc* analysis, Table S10). **(C)** (left panel) Mean normalized frequency response functions of responsive neurons tracked pre- to post-DFC across all *pseudo*-conditioned mice (N = 11 mice, N = 712 neurons). (right panel) Mean change in normalized frequency response functions for the same cells, square indicate significant changes (Two-way rm-ANOVA, *Tukey-Kramer post-hoc* analysis, Table S11).

Previous studies found that the best frequency of neurons shifts towards the conditioned stimulus (CS+) after DFC with pure tones [17]. We found no difference in best frequency changes between conditioned and *pseudo*-conditioned mice (Fig. 6A & B, two-way rm ANOVA, *Tukey-Kramer post-hoc* comparisons , p = .905, Table S20). However, consistent with a shift in best frequency towards the CS+, we did observe a small decrease in the absolute distance of best frequency from CS+ (of mean response functions pre- and post-DFC for each neuron) in responsive neurons of conditioned mice (Fig. 6C, -0.06 octaves, two-way rm ANOVA, *Tukey Kramer post-hoc*, p = .002, Table S21) but not in *pseudo*-conditioned mice (p = .969). It is possible that neuronal discrimination between CS+ and CS- could be altered by a change frequency tuning width [12]. As a measure of tuning width we used the sparseness of the frequency response function [29,30]: A neuron with high sparseness responds strongly to one or few frequencies tested and little to other frequencies. A neuron with a sparseness of zero would indicate an equal response to all frequencies tested. We found no difference between the conditoned and *pseudo*-conditioned mice (Fig. 6D, two-way rm-ANOVA, *Tukey-Kramer post-hoc* comparison $F_{(1,1883)} = 0.26$, $p = .610$, Table S22) and that sparseness decreased in both ($F_{(1,1883)} = 18.25$, $p < .001$).

In summary, we observed heterogeneous changes in responses of individual neurons tracked from pre- to post-DFC. In conditioned animals, there was, on average, an increase in normalized response at CS+ and no change at CS-, however increase was not observed in absolute response changes. In *pseudo*-conditioned mice, we observed no changes in normalized responses at the CS stimuli. We observed a small shift in best frequency towards CS+ in conditioned mice. Sparseness of the response functions decreased in both conditioned and *pseudo*-conditioned mice, indicating that frequency tuning became broader after conditioning, thus unlikely to support increased discriminability. Combined, these results reconcile our findings with previous studies, which had effectively, by not sampling responses from the same neurons pre- and post-DFC, normalized the responses. It is plausible that previous studies observed an increase in normalized activity, which did not translate into an actual population-wide increase in discriminability as we find here.

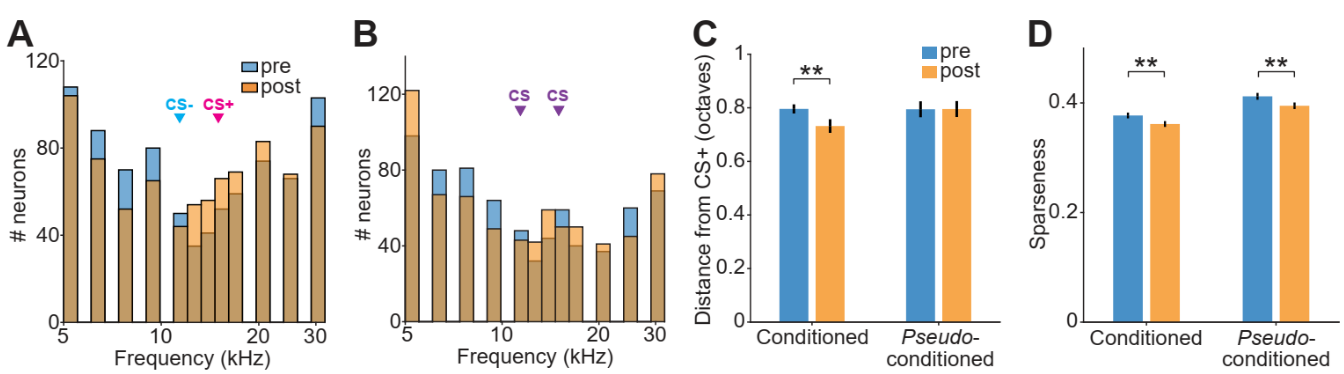

Figure 6: Best frequency and tuning sparseness pre- and post-conditioning. (A) Distributions of best frequencies of responsive neurons pre- (blue) and post-conditioning (orange), N = 826. **(B)** Same as **A** for *pseudo*-conditioned mice, N = 712. **(C)** Distance of best frequency from CS+ (15 kHz) of neurons from conditioned and *pseudo*-conditioned mice pre-

(blue) and post-conditioning (orange). Stats: 2-way ANOVA *Tukey-Kramer post-hoc* analysis (Table S21). **(D)** Sparseness of mean frequency response functions of neurons from conditioned and pseudo-conditioned mice pre- (blue) and post-conditioning (orange). Statistics: 2-way ANOVA *Tukey-Kramer post-hoc* analysis (Table S22). ** $p < .01$. Error bars = SEM.

A learning model of the fear circuit

We found that AC activity prior to learning predicts specificity of learning, yet the reorganized neuronal responses do not correlate with learning specificity. 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 neuronal 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 from previous work, in particular: (1) Discriminability between CS+ and CS- in AC predicts learning specificity post-DFC (Fig. 2-3); (2) Suppressing inhibition in AC leads to increased generalization (decreased learning specificity) post-DFC [12]; (3) Suppressing AC post-DFC does not affect learning specificity [3,11].

In the model, we included two populations of frequency-tuned neurons (representing the medial geniculate body, MGB, and AC). 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, the 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 to represent frequency discriminability (more overlap = less discriminability). The activity of the BLA after weight update and with auditory input only is used as a measure of freezing.

We first tested whether broad tuning in AC (low neuronal discriminability between CS+ and CS-) produced more generalized freezing than sharp tuning (high neuronal discriminability). We found that increased overlap in frequency tuning in AC neurons, without changing the tuning of MGB neurons, drove more generalized freezing responses (Fig. 7B, S9). This is due to the fact that, when AC was broadly tuned, CS+ tone activated AC neurons not only responded to the CS+ frequency but also to other frequencies, such as the 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. S9) 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. 2, 3). Second, we examined the effects of decreasing inhibition in the AC population during conditioning (Fig. 7C, S10). Decreasing inhibition resulted in an increased overlap in frequency responses in the AC population, which in turn led to increased generalization, supporting previous findings and providing a mechanism [12,31]. Third, we tested whether inactivating the auditory cortex following conditioning had an effect on freezing responses (Fig. 7C, S11). Consistent with previous findings [3,11], 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 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 identified here and in previous studies.

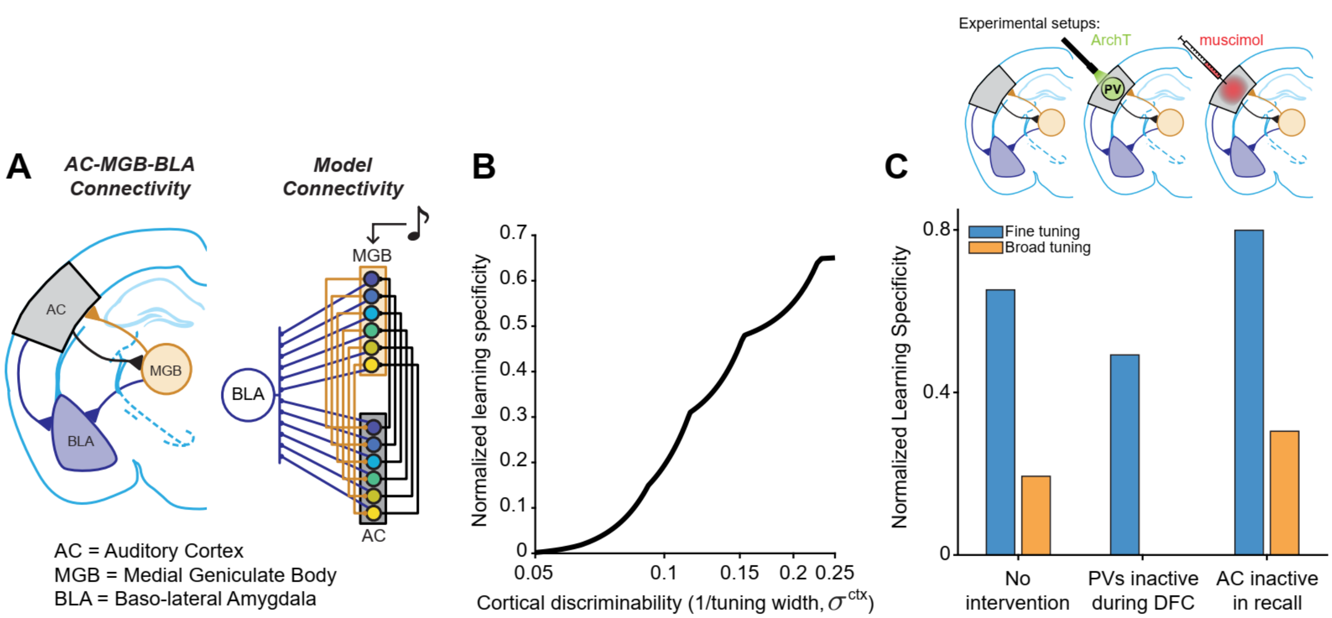

Figure 7: A learning model reconciles present and past findings. (A) (left panel) connectivity between auditory cortex (AC, gray), medial geniculate body (MGB, orange) and basolateral amygdala (BLA, blue). Right panel shows the model connectivity. MGB receives auditory input and provides input to AC (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 between the neurons in the AC population, σ^{ctx}. **(C)** Normalized learning specificity at two AC discriminability levels; fine (blue) and broad (orange) tuning. 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

Our results identify the role of the auditory cortex in differential fear learning: (1) Prior to fear learning, neuronal responses in AC shape fear learning specificity (Fig. 2 & 3); (2) Following differential fear conditioning, neuronal response transformations are not correlated with fear learning specificity (Fig. 5, Fig. S7), and therefore the auditory cortex does not encode auditory differential fear memory; (3) Neuronal activity in AC post-DFC does not correlate with freezing behavior (Fig. 4); (4) A simple model of the auditory nuclei and the basolateral amygdala could account for our results as well as a number of previous findings (Fig. 7).

Our finding that the neuronal activity prior to fear conditioning predicted specialization of fear learning provides a mechanism for the role of AC in differential fear memory acquisition [10–12,14,31]. Specifically, inactivation of inhibitory neurons in the AC during fear conditioning led to increased generalization of fear learning with pure tones [12]. Suppressing inhibitory neurons in the AC led to a decrease in Fisher information, which reflects the certainty about a stimulus in neuronal representation[31]. This change would likely result in a decrease in neuronal discriminability between the dangerous and safe tones in the AC, and therefore drive a increase in fear generalization, as demonstrated by our model (Fig. 7). Our results provide the link between optogenetic inactivation of interneurons in AC leading to increased fear generalization, and to increased frequency tuning width [12], which decreases neuronal discriminability.

By using two-photon imaging to record from the same neurons over the course of differential fear conditioning, we were able to compute changes in both absolute and relative neuronal activity of a large number of identified neurons, a feat not normally achievable with electrophysiology [17,19]. Previous work found that changes in neuronal responses to the dangerous and safe stimuli after differential fear conditioning amplified the difference between the responses [17,19]. This change was proposed to represent fear memory [15,19,28]. We identified similar transformations in the *normalized* response functions of neurons that were tracked pre- to post-conditioning, we found an increased relative response to the CS+. However, these changes did not correlate with freezing behavior suggesting that the neuronal code in the AC after fear conditioning does not reflect differential fear memory. Indeed, a number of studies found that inactivating the auditory cortex *after* fear conditioning with pure tones does not affect fear memory retrieval [3,11] (but see [14]). Combined, our results restrict the role of auditory cortex in fear conditioning to pure tone differential fear memory acquisition, but not retrieval.

If the increase in normalized response at CS+ is not related to fear memory, then why is there an increase in response? It could be reflective of increased attention caused by presentation of the CS+ and that the discrimination of the CS stimuli is unaffected by this effect [32]. Furthermore, changes in frequency map organization do not necessarily relate to changes in behavioral frequency discrimination of pure tones [33,34], thus over-representation of the CS+ could be induced by learning but not necessary for discrimination learning.

To locate our findings with previous work, we implemented a simple, anatomically accurate[35,36] model with connections from auditory nuclei to the basolateral amygdala (Fig. 7). The model demonstrated that (1) neuronal activity in cortex can predict subsequent learning specificity; that (2) inactivation of PV interneurons in AC during DFC leads to increased generalization [12], and that (3) the auditory cortex is not necessary for differential fear memory retrieval [3,11]. 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 pure tone stimuli during DFC by controlling stimulus discrimination in the MGB, this may or may not be a direct projection neuroanatomically[35,37,38]. Future studies need to explore the role of the MGB and specific projections between AC, MGB and BLA in fear learning and memory. It is likely that such an important behavioral modification as fear has redundant pathways to obtain the same behavioral outcomes[11,39–41].

Our results relied on tracking the neuronal responses in all transfected neurons in AC without distinguishing between different neuronal subtypes. Previous studies found that a specific class of inhibitory neuron increases activity with presentation of repeated tones [18,26]. It is therefore plausible that our results include a subset of neurons that function differently during fear conditioning but which we are unable to identify due to lack of selective labelling. Furthermore, we restricted our recordings to 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 [42,43]. We imaged neuronal activity in layers 2/3 because those are the output layers of the cortex, and therefore the plastic changes that occur within the cortex should be evident in these layers. The complexity of transformations in the cortical microcircuit and between layers with learning can be explored further [44–46].

The results of the study may be restricted to pure tone stimuli. We chose pure tone stimuli because these stimuli provide a well-defined axis (frequency) along which to vary stimulus discriminability. Furthermore, in human subjects, AC encodes threat during DFC for pure tone stimuli [21]. Our prior work has established that large frequency separation between CS+ and CS- results in uniform specificity of the fear response among subjects, whereas smaller frequency separation, such as the one used here, provides for a gradient of specificity across subjects [3,12]. Other studies have found that AC is not behaviorally relevant for discrimination between pure tones separated by large frequency distances[11,13]. However, when the frequencies were brought closer together, then manipulation of AC activity did affect behavior [13]. Therefore, it is unclear whether recent conclusions that AC is involved in processing of more complex stimuli and not pure tones are due to differences in complexity of the stimulus, or to the degree to which AC can discriminate these stimuli. Furthermore, the FM sweeps used in these studies are not necessarily more complex than pure-tones for AC processing. Indeed, neurons in the inferior colliculus, which is two synapses earlier than AC, differentiate between FM sweeps [e.g. 47]. Ultimately, the relevant aspect of the present study was the ability to measure how well neuronal ensembles differentiate between two stimuli. We achieved this by bringing CS+ and CS- close together in frequency, and we found that neuronal discriminability of the stimuli differs across mice and correlates with behavioral discriminability prior to DFC. We would not expect this result were the stimuli not relevant for AC. Future studies will dissect to what extent the differences in neuronal codes in AC shape differential fear learning of more complex and natural sounds and its role in other forms of learning [13,34,48,49].

Our results may be applicable to understanding anxiety disorders. An extreme example of fear generalization is realized in PTSD [50]. Here we find that the present state of each individual brain, in terms of neuronal 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 neuronal discrimination of potentially threatening stimuli [51–54]. Further work in this area can lead to a deeper understanding how genetic and social factors, as well early life experiences, shape the role of sensory cortex in this common and devastating disorder [7,53].

We identified a neuronal correlate for inter-individual differences in learning specificity. 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-amygdala 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 other 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., C.C, 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.

Online 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 (22 male, 10 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 (26) [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. 21 mice (15 males, 6 females) were in the conditioning group and 11 mice (7 males, 4 females) were in the *pseudo*-conditioning control group.

The Auditory Brainstem Response (ABR) to broadband clicks (2 – 80 kHz, 70 dB SPL) was acquired before or at the end of the experiment when possible in order to confirm that the mice had good hearing (Fig. S12).

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, 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 a pump (Pump 11 Elite, Harvard Apparatus, USA) and needles were pulled (P-97 Puller, Sutter Instruments, USA) glass pipettes (Harvard Apparatus, USA) with tip openings of 30 – 50 μm. 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 (Fig. 1). 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, 0.7 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). Fear memory retrieval sessions in context B followed each two-photon imaging session after conditioning. The CS+ and CS- were presented 4 times (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. Baseline freezing consisted of an equal time of silence prior to tone onset.

Conditioned mice that did not freeze either to CS+ or CS- were removed from subsequent analysis (two-way ANOVA for each mouse on freezing scores to CS+, CS- and baseline from all retrieval sessions (16 trials for each CS and 32 trials for baseline). Stimulus (CS+/CS-) and baseline (stimulus/no stimulus) were the independent variables. Learners were defined as those with significant effect of baseline or baseline*stimulus, $p < .05$). 6 mice (5 males, 1 female) were excluded from the study, leaving 15 conditioned mice (10 males, 5 females).

For each mouse the learning specificity (*LS*, Equation 1 [3]) was calculated as:

625 $$LS = \sum_{i=1}^{N} fr_{CS^+}(i) \Big/ N - \sum_{i=1}^{N} fr_{CS^-}(i) \Big/ N$$

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). We used a 16X Nikon objective with 0.8 numerical aperture (Thorlabs, N16XLWD-PF). 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 presented at 60 – 80 dB SPL. In a single recording session, each frequency was repeated 15 – 30 times in a *pseudo*-random order with a 4-s inter-stimulus interval.

Cell tracking across imaging sessions.

We imaged the activity from the same cells over 15 days in layers 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 (>2 standard deviations). ROIs which were not matched to any earlier ROIs were counted as new cells. 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 Fig. S1.

Data analysis and statistical procedures

Publicly available toolboxes [22] 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. This trace was low pass filtered (filter cut off at 7.5 Hz) to remove high frequency noise. From this filtered trace, we calculated the mean baseline fluorescence ($F_{baseline}$) and standard deviation of the baseline (F_{std}) over the one second prior to tone onset, and then determined the change in fluorescence over time relative to the mean baseline fluorescence ($\Delta F = F - F_{baseline}$) for each sound presentation. We then divided ΔF by F_{std}, effectively calculating the z-score of the fluorescence response relative to the baseline ($\Delta F/F_{std}$) for each sound presentation.

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 [55,56]). The frequency response function was defined as the mean response to each tone frequency across repeats. Best frequency was defined as the frequency with the highest mean response. Sparseness (S, Equation 2 [29,30]) was used to estimate the sharpness of response functions, with 1 being very sharply tuned and 0 being 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 (mean $\Delta F/F_{std}$ over 2 s post-stimulus onset) 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 with shuffled the CS+/CS- response labels 250 times. For mice not tested under the two-photon directly with the CS+ or CS-, the data were linearly interpolated to estimate responses at CS- and CS+.

For fitting the Support Vector Machine, we used MATLAB's *fitcsvm* function with a Gaussian kernel and 10-fold cross-validation to predict the learning specificity based on the standardized single-trial population responses (mean $\Delta F/F_{std}$ over 2 s post-stimulus onset for each neuron). We used a Gaussian kernel since it makes the least assumptions about the underlying distributions of population responses.

We calculated the confidence intervals of correlations using a bootstrap procedure, resampling, with replacement, the data 1000 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 bootstrapped r distributions of each dataset from one another, the change in r was considered significant if 95% CI of the difference-distribution did not overlap with zero.

To compare results between testing groups (conditioned/*pseudo*-conditioned) we used two-way repeated measures ANOVAs with the relevant variables (see Tables S1, S3-13, S16, S19-22,).

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 response functions pre- and post-DFC. For cells present in more than one session either pre- or post-DFC, the frequency response curves from each session were averaged and the changes in response were assessed from the mean across pre- and post-DFC sessions. For comparing the fluorescence traces of responses (Fig. S6D-G), for the 4 mice not tested directly at CS+ and CS-, the nearest frequencies were used.

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 and injected the red marker into the imaging site (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 before slicing. The location of imaging was confirmed through fluorescent imaging (Fig. S2). 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 [57].

Model

Neuron model and network

We simulated cortical neuronal populations, MGB populations and a BLA neuronal population in a rate-based description of neuronal activity. We simulated $N = 10$ MGB populations. Each MGB population receives $N = 10$ 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^{cxt} 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 neuronal populations always have a small overlap in neuronal responses (σ^{MGB} = 0.8).

Similarly, we simulated $N = 10$ cortical populations 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* 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 (σ^{ctx} = 3), and one corresponding to broad tuning with a large overlap (σ^{ctx} = 10). (Note that σ^{MGB} = 0.8 was equivalent to σ^{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: $\Delta {w^{ctx/MGB}}_i = \alpha {x^{ctx/MGB}}_i e$, 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 [12], 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 $x_i^{ctx} = 0$ during the BLA simulation protocol.

Data availability

All data and the code to generate the figures as well as the model are available in free access here: https://doi.org/10.5061/dryad.wpzgmsbhw .

REVIEWER LINK:

https://datadryad.org/stash/share/Cjd1A7BkhvPg2aNETgh3PAoAayMgRT0vIACCtSpBcTA

Supplemental Information (12 figures and 22 tables)

Table S3: Statistics for Fig. 4A

Table S4: Statistics for Fig. 4B

Table S5: Statistics comparing change in SVM performance over consecutive sessions.

Table S6: Statistics comparing change in Zdiff over consecutive sessions.

Table S7: Statistics for Fig. 5A (left panel).

Table S8: Statistics for Fig. 5A (middle panel).

Table S9: Statistics for Fig. 5A (right panel).

Table S10: Statistics for Fig. 5B Table

S11: Statistics for Fig. 5C

Table S12: Statistics comparing normalized responses to conditioned stimuli in conditioned and pseudo-conditioned mice from pre- to post-DFC.

Table S13: Statistics for Fig. S6B:

Table S14: Change in negative responses of the lower quartile of responding neurons from pre- to post-DFC in conditioned mice (Fig. S6E).

Table S15: Change in negative responses of the upper quartile of responding neurons from pre- to post-DFC in conditioned mice (Fig. S6E).

Table S16: Statistics for Fig. S6C.

Table S17: Change in negative responses of the lower quartile of responding neurons from pre- to post-DFC in pseudo-conditioned mice (Fig. S6G).

Table S18: Change in negative responses of the upper quartile of responding neurons from pre- to post-DFC in pseudo-conditioned mice (Fig. S6G).

Table S19: Statistics comparing non-normalized responses to conditioned stimuli in conditioned and pseudo-conditioned mice from pre- to post-DFC.

Table S20: Statistics for Fig. 6A & B.

Table S21: Statistics for Fig. 6C.

Table S22: Statistics for Fig. 6D.

References

1. Quirk, G. J., Armony, J. L. & LeDoux, J. E. Fear conditioning enhances different temporal components of tone-evoked spike trains in auditory cortex and lateral amygdala. *Neuron* **19**, 613–624 (1997).
2. Johansen, J. P., Cain, C. K., Ostroff, L. E. & LeDoux, J. E. Molecular Mechanisms of Fear Learning and Memory. *Cell* **147**, 509–524 (2011).
3. Aizenberg, M. & Geffen, M. N. Bidirectional effects of aversive learning on perceptual acuity are mediated by the sensory cortex. *Nat. Neurosci.* **16**, 994–6 (2013).
4. Resnik, J., Sobel, N. & Paz, R. Auditory aversive learning increases discrimination thresholds. *Nat. Neurosci.* **14**, 791–796 (2011).
5. Li, W., Howard, J. D., Parrish, T. B. & Gottfried, J. A. Aversive learning enhances perceptual and cortical discrimination of indiscriminable odor cues. *Science (80-.).* **319**, 1842–1845 (2008).
6. Chapuis, J. & Wilson, D. a. Bidirectional plasticity of cortical pattern recognition and behavioral sensory acuity. *Nature Neuroscience* **15**, 155–161 (2011).
7. Mahan, A. L. & Ressler, K. J. Fear conditioning, synaptic plasticity and the amygdala: Implications for posttraumatic stress disorder. *Trends in Neurosciences* **35**, 24–35 (2012).
8. Jovanovic, T. & Ressler, K. J. How the neurocircuitry and genetics of fear inhibition may inform our understanding of PTSD. *American Journal of Psychiatry* **167**, 648–662 (2010).
9. Krusemark, E. A. & Li, W. Enhanced olfactory sensory perception of threat in anxiety: An event-related fMRI study. *Chemosens. Percept.* **5**, 37–45 (2012).
10. Letzkus, J. *et al.* A disinhibitory microcircuit for associative fear learning in the auditory

cortex. *Nature* **480**, 331–336 (2011).

11. Dalmay, T. *et al.* A Critical Role for Neocortical Processing of Threat Memory. *Neuron* 1–15 (2019). doi:10.1016/J.NEURON.2019.09.025

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

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

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

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

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

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

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

19. Ohl, F. W. & Scheich, H. Differential frequency conditioning enhances spectral contrast sensitivity of units in auditory cortex (field AI) of the alert Mongolian gerbil. *Eur. J. Neurosci.*

8, 1001–17 (1996).

20. Chen, T.-W. *et al.* Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499**, 295–300 (2013).
21. Staib, M. & Bach, D. R. Stimulus-invariant auditory cortex threat encoding during fear conditioning with simple and complex sounds. *Neuroimage* **166**, 276–284 (2018).
22. Pachitariu, M. *et al. Suite2p: beyond 10,000 neurons with standard two-photon microscopy*. *bioRxiv* **53**, (Cold Spring Harbor Labs Journals, 2017).
23. Georgopoulos, A. P., Schwartz, A. B. & Kettner, R. E. Neuronal Population Coding of Movement Direction. *Science* **233**, 1416–1419 (1986).
24. Stringer, C. *et al.* Spontaneous behaviors drive multidimensional, brainwide activity. *Science (80-.).* **364**, (2019).
25. Runyan, C. A., Piasini, E., Panzeri, S. & Harvey, C. D. Distinct timescales of population coding across cortex. *Nature* **548**, 92–96 (2017).
26. Kato, H. K., Gillet, S. N. & Isaacson, J. S. Flexible Sensory Representations in Auditory Cortex Driven by Behavioral Relevance. *Neuron* **88**, 1027–1039 (2015).
27. Ohl, F. W. & Scheich, H. Learning-induced dynamic receptive field changes in primary auditory cortex of the unanaesthetized Mongolian gerbil. *J. Comp. Physiol. A Sensory, Neural, Behav. Physiol.* **181**, 685–696 (1997).
28. Ohl, F. W. & Scheich, H. Fallacies in behavioural interpretation of auditory cortex plasticity. *Nat. Rev. Neurosci.* **5**, 1–1 (2004).
29. Rust, N. C. & DiCarlo, J. J. Balanced Increases in Selectivity and Tolerance Produce Constant Sparseness along the Ventral Visual Stream. *J. Neurosci.* **32**, 10170–10182

(2012).

30. Blackwell, J. M., Lesicko, A. M., Rao, W., De Biasi, M. & Geffen, M. N. Auditory cortex shapes sound responses in the inferior colliculus. *Elife* **9**, (2020).

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

32. Pooresmaeili, A., Poort, J., Thiele, A. & Roelfsema, P. R. Separable Codes for Attention and Luminance Contrast in the Primary Visual Cortex. *J. Neurosci.* **30**, 12701–12711 (2010).

33. Talwar, S. K. & Gerstein, G. L. Reorganization in awake rat auditory cortex by local microstimulation and its effect on frequency-discrimination behavior. *J. Neurophysiol.* **86**, 1555–1572 (2001).

34. Maor, I. *et al.* Neural Correlates of Learning Pure Tones or Natural Sounds in the Auditory Cortex. *Front. Neural Circuits* **13**, 82 (2020).

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

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

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

38. Aizenberg, M. *et al.* Projection from the Amygdala to the Thalamic Reticular Nucleus Amplifies Cortical Sound Responses. *Cell Rep.* **28**, 605-615.e4 (2019).

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

40. Zhang, G.-W. *et al.* A Non-canonical Reticular-Limbic Central Auditory Pathway via Medial Septum Contributes to Fear Conditioning. *Neuron* **97**, 406-417.e4 (2018).

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

42. Atencio, C., Sharpee, T. & Schreiner, C. Receptive field dimensionality increases from the auditory midbrain to cortex. *J. Neurophysiol.* (2012). doi:10.1152/jn.01025.2011

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

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

45. Wood, K. C., Blackwell, J. M. & Geffen, M. N. Cortical inhibitory interneurons control sensory processing. *Current Opinion in Neurobiology* **46**, 200–207 (2017).

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

47. Kuo, R. I. & Wu, G. K. The Generation of Direction Selectivity in the Auditory System. *Neuron* **73**, 1016–1027 (2012).

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

49. Betzel, R. F., Wood, K. C., Angeloni, C., Neimark Geffen, M. & Bassett, D. S. Stability of spontaneous, correlated activity in mouse auditory cortex. *PLOS Comput. Biol.* **15**,

e1007360 (2019).

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

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

52. Lange, I. *et al.* Behavioral pattern separation and its link to the neural mechanisms of fear generalization. *Soc. Cogn. Affect. Neurosci.* **12**, 1720–1729 (2017).

53. Roesmann, K. *et al.* Fear generalization of implicit conditioned facial features – Behavioral and magnetoencephalographic correlates. *Neuroimage* **205**, 116302 (2020).

54. Tuominen, L. *et al.* The relationship of perceptual discrimination to neural mechanisms of fear generalization. *Neuroimage* **188**, 445–455 (2019).

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

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

57. Shamash, P., Carandini, M., Harris, K. & Steinmetz, N. A tool for analyzing electrode tracks from slice histology. *bioRxiv* 447995 (2018). doi:10.1101/447995